# ROBUST AGENTS LEARN CAUSAL WORLD MODELS

**Jonathan Richens***
Google DeepMind

**Tom Everitt**
Google DeepMind

## ABSTRACT

It has long been hypothesised that causal reasoning plays a fundamental role in robust and general intelligence. However, it is not known if agents must learn causal models in order to generalise to new domains, or if other inductive biases are sufficient. We answer this question, showing that any agent capable of satisfying a regret bound under a large set of distributional shifts must have learned an approximate causal model of the data generating process, which converges to the true causal model for optimal agents. We discuss the implications of this result for several research areas including transfer learning and causal inference.

## 1 INTRODUCTION

What capabilities are necessary for general intelligence (Legg & Hutter, 2007)? One candidate is causal reasoning, which plays a foundational role in human cognition (Gopnik et al., 2007; Sloman & Lagnado, 2015). It has even been argued that human-level AI is impossible without causal reasoning (Pearl, 2018). However, recent years have seen the development of agents that do not explicitly learn or reason on causal models, but nonetheless are capable of adapting to a wide range of environments and tasks (Reed et al., 2022; Team et al., 2023; Brown et al., 2020).

This raises the question, do agents have to learn causal models in order to adapt to new domains, or are other inductive biases sufficient? To answer this question, we have to be careful not to assume that agents use causal assumptions a priori. For example, transportability theory determines what causal knowledge is necessary for transfer learning when all assumptions on the data generating process (inductive biases) can be expressed as constraints on causal structure (Bareinboim & Pearl, 2016). However, deep learning algorithms can exploit a much larger set of inductive biases (Neyshabur et al., 2014; Battaglia et al., 2018; Rahaman et al., 2019; Goyal & Bengio, 2022) which in many real-world tasks may be sufficient to identify low regret policies without requiring causal knowledge.

The main result of this paper is to answer this question by showing that,

*Any agent capable of adapting to a sufficiently large set of distributional shifts must have learned a causal model of the data generating process.*

Here, adapting to a distributional shift means learning a policy that satisfies a regret bound following an intervention on the data generating process—for example, changing the distribution of features or latent variables. On the one hand, a causal model of the data generating process can be used to identify regret-bounded policies following a distributional shift (sufficiency), with more accurate causal models allowing lower regret policies to be identified. We prove the converse (necessity)—given regret-bounded policies for a large set of distributional shifts, we can learn an approximate causal model of the data generating process, with the approximation becoming exact for optimal policies. Hence, learning a causal model of the data generating process is both necessary and sufficient for robust adaptation.

This equivalence has consequences for a number of fields and questions. For one, it implies that an agent's ability to adapt to distributional shifts is bounded by its ability to learn causal relations between environment variables, opening a path to deriving fundamental causal limitations for transfer learning using impossibility results for causal discovery. For example, we show that adapting to covariate and label shifts in simple supervised learning tasks is only possible if the causal relations between features and labels can be identified from the training data—a non-trivial causal discovery

*jonrichens@deepmind.com

problem. Our result also adds justification to causal representation learning (Schölkopf et al., 2021), by showing that learning causal representations is necessary for achieving robustness across many domains. More speculatively, it suggests that causal models may play a role in emergent capabilities, as agents trained to optimise a single objective function across many domains are forced to learn a causal model which in turn enables them (in principle) to reason about interventions, to counterfact, and to optimise a much larger set of objective functions.

**Outline of paper.** In Section 2 we introduce concepts from causality and decision theory used to derive our results. We present our main theoretical results in Section 3 and discuss their interpretation in terms of adaptive agents, transfer learning and causal inference. In Section 4 we discuss limitations, as well as implications for a number of fields and open questions. In section 5 we discuss related work including transportability (Bareinboim & Pearl, 2016) and the causal hierarchy theorem (Bareinboim et al., 2022), and recent empirical work on emergent world models. In Appendix B we describe experiments applying our theoretical results to causal discovery problems.

## 2 PRELIMINARIES

### 2.1 CAUSAL MODELS

We use capital letters for random variables $V$, and lower case for their values $v \in dom(V)$. For simplicity, we assume each variable has a finite number of possible values, $|dom(V)| < \infty$. Bold face denotes sets of variables $\boldsymbol{V} = \{V_1, \ldots, V_n\}$, and their values $\boldsymbol{v} \in dom(\boldsymbol{V}) = \times_i dom(V_i)$. A probabilistic model specifies the joint distribution $P(\boldsymbol{V})$ over a set of variables $\boldsymbol{V}$. These models can support associative or observational queries, for example $P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$. Interventions describe external changes to the data generating process (and hence changing the joint distribution), for example a *hard* intervention $\text{do}(\boldsymbol{X} = \boldsymbol{x})$ describes forcing the set of variables $\boldsymbol{X} \subseteq \boldsymbol{V}$ to take value $\boldsymbol{x}$. This generates a new distribution $P(\boldsymbol{V} \mid \text{do}(X = x)) = P(\boldsymbol{V}_x)$ where $\boldsymbol{V}_{\boldsymbol{x}}$ refers to the variables $\boldsymbol{V}$ following this intervention. The power of causal models is that they specify not only $P(\boldsymbol{V})$ but also the distribution of $\boldsymbol{V}$ under all interventions, and hence these models can be used to evaluate both associative and interventional queries e.g. $P(\boldsymbol{Y} = \boldsymbol{y} \mid \text{do}(\boldsymbol{X} = \boldsymbol{x}))$.

For the derivation of our results we focus on a specific class of causal models—causal Bayesian networks (CBNs). There are several alternative models and formalisms that are studied in the literature, including structural equation models (Pearl, 2009) and the Neyman-Rubin causal models (Rubin, 2005), and results can be straightforwardly adapted to these.

**Definition 1** (Bayesian networks)**.** *A* Bayesian network $M = (G, P)$ *over a set of variables* $\boldsymbol{V} = \{V_1, \ldots, V_n\}$ *is a joint probability distribution* $P(\boldsymbol{V})$ *that factors according to a directed acyclic graph (DAG)* $G$*, i.e.* $P(V_1, \ldots, V_n) = \prod_{i=1}^{n} P(V_i \mid \boldsymbol{Pa}_{V_i})$*, where* $\boldsymbol{Pa}_{V_i}$ *are the parents of* $V_i$ *in* $G$*.*

A Bayesian network is *causal* if the graph $G$ captures the causal relationships between the variables or, formally, if the result of any intervention $\text{do}(\boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{X} \subseteq \boldsymbol{V}$ can be computed from the truncated factorisation formula:

$$P(\boldsymbol{v} \mid \text{do}(\boldsymbol{x})) = \begin{cases} \prod_{i:v_i \notin \boldsymbol{x}} P(v_i \mid \mathbf{pa}_{v_i}) & \text{if } \boldsymbol{v} \text{ consistent with } \boldsymbol{x} \\ 0 & \text{otherwise.} \end{cases}$$

More generally, a *soft* intervention $\sigma_{v_i} = P'(V_i \mid \mathbf{Pa}_i^*)$ replaces the conditional probability distribution for $V_i$ with a new distribution $P'(V_i \mid \mathbf{Pa}_i^*)$, possibly resulting in a new parent set $\mathbf{Pa}_i^* \neq \mathbf{Pa}_i$ as long as no cycles are introduced in the graph. We refer to $\sigma_{v_i}$ as a *domain indicator* (Correa & Bareinboim, 2020) (it has also been called an environment index, Arjovsky et al., 2019). The updated distribution is denoted $P(\boldsymbol{v}; \sigma_{\boldsymbol{v}'}) = \prod_{i:v_i \in \boldsymbol{v}'} P'(v_i \mid \mathbf{pa}_{v_i}^*) \prod_{i:v_i \notin \boldsymbol{v}'} P(v_i \mid \mathbf{pa}_{v_i})$.

In general, soft interventions cannot be defined without knowledge of $G$. For example, the soft intervention $\sigma_Y = P'(y \mid x)$ is incompatible with the causal structure $Y \to X$ as it would induce a causal cycle. As our results are concerned with learning causal models (and hence causal structure), we focus our theoretical analysis on a subset of the soft interventions, *local interventions*, that are compatible with all causal structures and so can be used without tacitly assuming knowledge of $G$.

**Definition 2** (Local interventions)**.** *Local intervention $\sigma$ on $V_i \in \boldsymbol{V}$ involves applying a map to the states of $V_i$ that is not conditional on any other endogenous variables, $v_i \mapsto f(v_i)$. We use the notation $\sigma = do(V_i = f(v_i))$ (variable $V_i$ is assigned the state $f(v_i)$). Formally, this is a soft intervention on $V_i$ that transforms the conditional probability distribution as,*

$$P(v_i \,|\, \boldsymbol{pa}_i; \sigma) = \sum_{v_i' : f(v_i') = v_i} P(v_i' \,|\, \boldsymbol{pa}_i) \tag{1}$$

*Example:* Hard interventions $\text{do}(V_i = v_i')$ are local interventions where $f(v_i)$ is a constant function.

*Example:* Translations are local interventions as $\text{do}(V_i = v_i + k) = \text{do}(V_i = f(v_i))$ where $f(v_i) = v_i + k$. Examples include changing the position of objects in RL environments (Shah et al., 2022) and images (Engstrom et al., 2019).

*Example:* Logical NOT operation $X \mapsto \neg X$ for Boolean $X$ is a local intervention.

We also consider stochastic interventions, noting that mixtures of local interventions can also be defined without knowledge of $G$. For example, adding noise to a variable $X = X + \epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$, is a soft intervention on $X$ described by a mixture over local interventions (translations).

**Definition 3** (Mixtures of interventions)**.** *A* mixed intervention $\sigma^* = \sum_i p_i \sigma_i$ *for* $\sum p_i = 1$ *performs intervention $\sigma_i$ with probability $p_i$. Formally, $P(\boldsymbol{v} \mid \sigma^*) = \sum_i p_i P(\boldsymbol{v} \mid \sigma_i)$.*

### 2.2 Decision tasks

Decision tasks involve a decision maker (agent) choosing a policy so as to optimise an objective function (utility). To give a causal description of decision tasks we use the causal influence diagram (CID) formalism (Howard & Matheson, 2005; Everitt et al., 2021), which extend a CBN of the environment (chance) variables by introducing decision and utility nodes (see Figure 1 for examples). For simplicity we focus on tasks involving a single decision and a single utility function.

**Definition 4** (Causal influence diagram)**.** *A (single-decision, single-utility)* causal influence diagram *(CID) is a CBN $M = (G, P)$ where the variables $\boldsymbol{V}$ are partitioned into decision, utility, and chance variables, $\boldsymbol{V} = (\{D\}, \{U\}, \boldsymbol{C})$. The utility variable is a real-valued function of its parents, $U(\boldsymbol{pa}_U)$.*

Single-decision single-utility CIDs can represent most decision tasks such as classification and regression as they specify what decision should be made ($d \in D$), based on what information ($\mathbf{pa}_D$), and towards what objective ($\mathbb{E}[U]$). They can also describe some multi-decision tasks such as Markov decision processes[1]. The utility can be any real-valued function and so can model any standard loss or reward functions.

We assume that the environment is described by a set of random variables $\boldsymbol{C}$ that interact via causal mechanisms[2], and where $\boldsymbol{C}$ satisfies causal sufficiency (Pearl, 2009) (includes all common causes), noting that such a choice of $\boldsymbol{C}$ always exists. We refer to the CBN over $\boldsymbol{C}$ as the 'true' or 'underlying' CBN. Note we do not assume the agent has any knowledge of the underling CBN, nor do we assume which variables in $\boldsymbol{C}$ are observed or unobserved by the agent, beyond that the agent observes its own inputs $\mathbf{Pa}_D \subseteq \boldsymbol{C}$. We also assume knowledge of the utility function $U(\mathbf{Pa}_U)$.

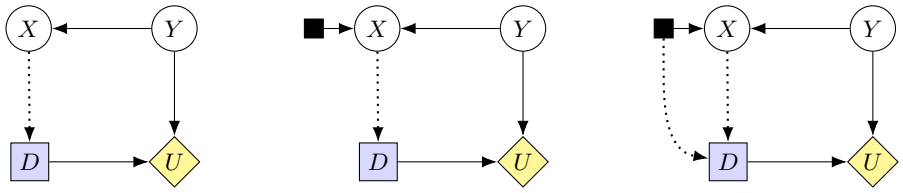
The conditional probability distribution for the decision node $\pi(d \mid \mathbf{pa}_D)$ (the policy) is not a fixed parameter of the model but is set by the agent so as to maximise its expected utility, which for a policy $\pi$ is $\mathbb{E}^\pi[U] = \mathbb{E}[U \mid \text{do}(D = \pi(\mathbf{pa}_D))]$. A policy $\pi^*$ is *optimal* if it maximises $\mathbb{E}^{\pi^*}[U]$. Typically, agents do not behave optimally and incur some *regret* $\delta$, which is the decrease in expected utility compared to an optimal policy $\delta := \mathbb{E}^{\pi^*}[U] - \mathbb{E}^\pi[U]$.

To simplify our theoretical analysis, we focus on a widely studied class of decision tasks where the agents policy is evaluated directly, i.e. where there are no mediators between the agent's decision $D$ and the utility $U$ (for examples see Figure 1).

**Assumption 1** (Unmediated decision task)**.** $\boldsymbol{Desc}_D \cap \boldsymbol{Anc}_U = \emptyset$.

[1] Note Markov decision processes can be formulated as a single-decision single-utility CID, by modelling the choice of policy as a single decision and the cumulative discounted reward as a single utility variable.

[2] This assumption follows from Reichenbach (1956), and we discuss further in Appendix A.3




(a) training (b) testing (c) known domain shift

Figure 1: CID for a supervised learning task during (a) training and (b) testing following a distributional (covariate) shift (unsupervised domain adaptation, Wilson & Cook, 2020). The agent chooses a label prediction $D = \hat{Y}$ given features $X$, with the goal of minimising loss $U = -\text{Loss}(Y, \hat{Y})$. Decision variables are depicted as square nodes, chance variables as circular nodes and utilities as diamond nodes. Information edges (dashed) show the variables the agent conditions their policy on. In this example the labels cause the features $Y \to X$ (for examples where features cause labels see Castro et al., 2020; Schölkopf et al., 2012). The black square ('regime node' (Correa & Bareinboim, 2020)) in (b) and (c) denotes a distributional shift induced by an intervention on $X$. Diagram (c) depicts the idealised case where the agent knows what domain shift has occurred. By theorem 1, if the agent can return an optimal decision boundary for known covariate and label shifts, then it must have learned the CBN over $\boldsymbol{C} = \{X, Y\}$. Note that even if the agent has sufficient training data to learn $P(X, Y)$, the causal structure $Y \to X$ is in general non-identifiable given $P(X, Y)$ and so domain adaptation requires that the agent solves a non-trivial causal discovery problem.

Assumption 1 implies that the agent's decision does not causally influence the environment except for the utility function. In unmediated decision tasks, the agent is provided some (partial) observations of the environment and chooses a policy, which is then evaluated using the utility function which is a function of the environment state and the agent's decision. Examples of unmediated decision tasks include prediction tasks such as classification and regression, whereas examples of *mediated* decision tasks that are not covered by our theorems include Markov decision processes where the agent's decision (action) influences the utility via the environment state.

### 2.3 Distributional shifts

We focus on generalisation that goes beyond the *iid* assumption, where agents are evaluated in domains that are *distributionally shifted* from the training environment (Farahani et al., 2021; Wilson & Cook, 2020; Shen et al., 2021; Zhou et al., 2022). Our analysis focuses on distributional shifts that involve changes to the causal data generating process, and hence can be modelled as interventions (Schölkopf et al., 2021). This does not assume that all shifts an agent will encounter can be modelled as interventions, but requires that the agent is *at least* capable of adapting to these shifts.

Examples of interventionally generated shifts include translating objects in images (Engstrom et al., 2019; Azulay & Weiss, 2018), noising inputs and adversarial robustness (Hendrycks & Dietterich, 2019), and changes to the initial conditions or transition function in Markov decision processes (Di Langosco et al., 2022; Eysenbach & Levine, 2021; Peng et al., 2018). Examples of shifts that are not naturally represented as interventions include changing the set of environment variables $\boldsymbol{C}$, and introducing selection biases (Shen et al., 2018). See Appendix A.3 for further discussion.

Distributional shifts can either describe changes to the environment, as in domain adaptation (Farahani et al., 2021; Wilson & Cook, 2020) and domain generalisation (Farahani et al., 2021), or changes to the objective, as in zero shot learning (Xian et al., 2018), in-context learning (Brown et al., 2020) and multi-task reinforcement learning (Reed et al., 2022). Our main results (Theorems 1 and 2) restrict to *domain shifts*, which correspond to interventions $\sigma$ on the chance variables $\boldsymbol{C}$. In contrast, *task shifts* change the objective (utility function), and are described by soft interventions on $\sigma_U$ replacing $U(\mathbf{Pa}_U)$ with a new utility function $U'(\mathbf{Pa}_{U'})$.

As we are interested in determining the capabilities necessary for domain adaptation, we restrict our attention to decision tasks where domain adaptation is non-trivial, i.e. where the optimal policy depends on the environment distribution $P(\boldsymbol{C} = \boldsymbol{c})$.

**Assumption 2** (Domain dependence)**.** *There exists $P(\boldsymbol{C} = \boldsymbol{c})$ and $P'(\boldsymbol{C} = \boldsymbol{c})$ compatible with $M$ such that $\pi^* = \arg\max_\pi \mathbb{E}^\pi_P[U]$ implies $\pi^* \neq \arg\max_\pi \mathbb{E}^\pi_{P'}[U]$.*

Assumption 2 implies the existence of domain shifts that change the optimal policy. As shown in Appendix A.1, domain independence holds if and only if $\mathbf{Anc}_U \subseteq \mathbf{Pa}_D$, i.e. when there are no latent variables (Lemma 1). It is simple to show that in this case the optimal policy is invariant under all domain shifts. Hence, Assumption 2 is equivalent to assuming the presence of latent variables that are strategically relevant (Koller & Milch, 2003).

## 3 Necessity of causal models for domain adaptation

We now present our main results, showing that learning the underlying CBN is necessary and sufficient for learning regret bounded policies under a large set of domain shifts. In Section 3.2 we interpret these theorems by considering their consequences for transfer learning, causal inference and adaptive agents.

First, we focus on the idealised case where we assume we have learned optimal policies for a large number of distributional shifts, and reconstruct the underlying CBN given these policies alone.

**Theorem 1.** *For almost all CIDs $M = (G, P)$ satisfying Assumptions 1 and 2, we can identify the directed acyclic graph $G$ and joint distribution $P$ over all ancestors of the utility $\boldsymbol{Anc}_U$ given $\{\pi^*_\sigma(d \mid \boldsymbol{pa}_D)\}_{\sigma\in\Sigma}$ where $\pi^*_\sigma(d \mid \boldsymbol{pa}_D)$ is an optimal policy in the domain $\sigma$ and $\Sigma$ is the set of all mixtures of local interventions. Proof in Appendix C.*

The parameters $P(v_i \mid \mathbf{pa}_i)$, $U(\mathbf{pa}_U)$ of the underlying CBN define a parameter space and the condition *for almost all CIDs* means that the subset of the parameter space for which the Theorem 1 does not hold is Lebesgue measure zero (see Appendix A.2 for discussion). This condition is necessary because there exist finely-tuned environments for which the CBN cannot be identified given the agent's policy due to variables $X \in \mathbf{Anc}_U$ that do not affect the expected utility. For example consider $X \to Y \to U$, $Y = \mathcal{N}(0, x)$ and $U = D + Y$, then changing $X$ can only change the variance of $U$ while leaving its expected value (and hence the optimal policy) constant. However, this only occurs for very specific choices of the parameters $P$ and $U$.

In Appendix B we give a simplified overview of the proof with a worked example. We assume access to an oracle for optimal policies $\pi^*_\sigma$ for any given local intervention $\sigma$. We devise an algorithm that queries this oracle with different mixtures of local interventions and identifies the mixtures for which the optimal policies changes. We then show that these critical mixtures identify the parameters of the CBN, specifying both the graph $G(\mathbf{Anc}_U)$ and the joint distribution $P(\mathbf{Anc}_U)$.

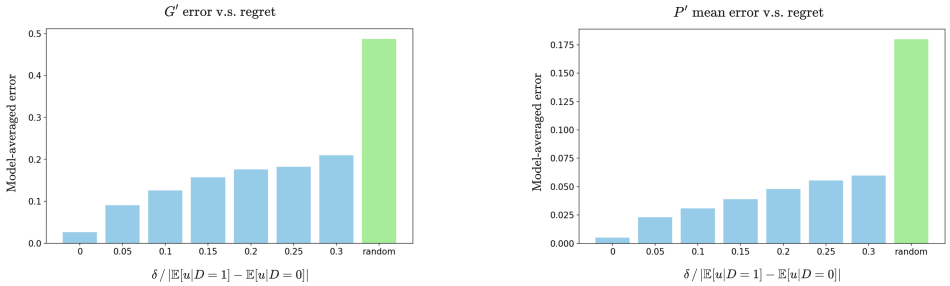
### 3.1 Relaxing the assumption of optimality

The strongest assumption in Theorem 1 is that we know optimal policies under domain shifts, whereas in most realistic settings the aim is to learn policies satisfying a regret bound (Mansour et al., 2009). Here we relax this assumption, considering the case where we know policies $\pi_\sigma$ that satisfy a given regret bound $\mathbb{E}^{\pi_\sigma}[U] \geq \mathbb{E}^{\pi^*_\sigma}[U] - \delta$. We show that for $\delta > 0$ it is possible to identify an approximation of the environment CBN, with error that grows linearly in $\delta$ for $\delta \ll \mathbb{E}^{\pi^*}[U]$.

**Theorem 2.** *For almost all CIDs $M = (G, P)$ satisfying Assumptions 1 and 2, we can identify an approximate causal model $M' = (P', G')$ given $\{\pi_\sigma(d \mid \boldsymbol{pa}_D)\}_{\sigma\in\Sigma}$ where $\mathbb{E}^{\pi_\sigma}[U] \geq \mathbb{E}^{\pi^*_\sigma}[U] - \delta$ and $\Sigma$ is the set of mixtures of local interventions. The parameters of $M'$ satisfy $|P'(v_i \mid \boldsymbol{pa}_i) - P(v_i \mid \boldsymbol{pa}_i)| \leq \gamma(\delta)\ \forall\ V_i \in \boldsymbol{V}$ where $\gamma(0) = 0$ and $\gamma(\delta)$ grows linearly in $\delta$ for small regret $\delta \ll \mathbb{E}^{\pi^*}[U]$. Proof in Appendix D.*

The worst-case bounds $\gamma(\delta)$ for the parameter errors are detailed in Appendix D. For $\delta > 0$ it may not be possible to identify $G$ perfectly due to the presence of weak causal relations cannot be resolved due to these error bounds. In this case we describe in Appendix D how we can learn a sub-graph $G' \subseteq G$ that includes the parent-child causal edges for which the causal relation is strong enough to be identified. Theorem 2 shows that we can learn a (sparse) approximate causal models of the data generating process from regret bounded policies under domain shifts, where the approximation becoming exact as $\delta \to 0$. In Appendix F we demonstrate learning the underlying CBN from

regret-bounded policies using simulated data for randomly generated CIDs where $\boldsymbol{C} = \{X, Y\}$ (e.g. Figure 1), and explore how the accuracy of the approximate CBN scales with the regret bound (partial results displayed in Figure 3). We now prove sufficiency, i.e. that having an (approximate) causal




(a) Error rate for learned DAG v.s. regret bound

(b) Mean error for $P(x, y)$ v.s. regret bound

Figure 3: Comparing the model-average error rates for a) the learned DAG $G'$ and b) learned joint distribution $P'(x, y)$, v.s. the (normalised) regret bound $\delta / \left|\mathbb{E}[u \mid D = 1] - \mathbb{E}[u \mid D = 0]\right|$. Average error taken over 1000 randomly generated environments with binary decision $D$ and two binary latent variables $X, Y$. Comparison to error rate for random guess (green) See Appendix F for details.

model of the data generating process is sufficient to identify regret-bounded policies. The result is well-known for the non-approximate case (Bareinboim & Pearl, 2016).

**Theorem 3.** *Given the CBN $M = (P, G)$ that is causally sufficient we can identify optimal policies $\pi^*_\sigma(d \mid \boldsymbol{pa}_D)$ for any given $U$ and for all soft interventions $\sigma$. Given an approximate causal model $M' = (P', G')$ for which $|P'(v_i \mid \boldsymbol{pa}_i) - P(v_i \mid \boldsymbol{pa}_i)| \leq \epsilon \ll 1$, we can identify regret-bounded policies where the regret $\delta$ grows linearly in $\epsilon$. Proof in Appendix E.*

Together, Theorems 2 and 3 imply that learning an approximate causal model of the data generating process is necessary and sufficient for learning regret-bounded policies under local interventions.

### 3.2 Interpretation

We interpret Theorems 1 to 3 through three lenses; transfer learning, adaptive agents and causal inference.

**Transfer learning.** In transfer learning (Zhuang et al., 2020), models are trained on a set of source domains and evaluated on held-out target domains where i) the data distribution differs from the source domains, and ii) the data available for training is restricted compared to the source domains (Wang et al., 2022). For example, in unsupervised domain adaptation the learner is restricted to samples of the input features from the target domains $\mathbf{pa}_D \sim P(\mathbf{Pa}_D; \sigma)$, whereas in domain generalisation typically no data from the target domain is available during training (Farahani et al., 2021).

Let $\mathcal{D}_S$ denote the training data from the source domains and $\mathcal{D}_\sigma$ denote the training data available from a given target domain $\sigma$. Let there exist a transfer learning algorithm that returns a policy $\pi_\sigma$ satisfying a regret bound for a given target domain $\sigma \in \Sigma$, given this training data. As $\pi_\sigma$ is a function of the training data, then by Theorems 1 and 2 the existence of this algorithm implies that we can identify the underlying CBN from $\mathcal{D}_S \cup \{\mathcal{D}_\sigma\}_{\sigma \in \Sigma}$. To see that this imparts non-trivial constraints on the existence of the transfer learning algorithm, we can consider the following simple example.

*Example:* Consider the CID for a simple supervised learning task with no latent confounders between features and labels, depicted in Figure 1. By Theorem 1, if there exists a transfer learning algorithm that identifies an optimal policy under all domain and label shifts, then the underlying CBN can be identified from $\mathcal{D}_S \cup \{\mathcal{D}_\sigma\}_{\sigma \in \Sigma}$. Let $\mathcal{D}_S = \{(x^i, y^i) \sim P(X, Y)\}_{i=1}^n$, so for sufficiently large $n$ the agent can learn the joint distribution $P(X, Y)$ from $\mathcal{D}_S$. However, the DAG $Y \to X$ must also be identifiable from the training data. In other words, the transfer learning problem contains a hidden causal discovery problem. If $\mathcal{D}_\sigma = \emptyset$ the agent has to identify this causal relation from $P(x, y)$ alone, which is impossible without additional assumptions (see for example Hoyer et al., 2008). If the agent has unlabelled features in the target domain $\mathcal{D}_\sigma = \{x^i \sim P(X; \sigma)\}_{i=1}^{n_\sigma}$ then we can in principle identify if $Y \to X$ by observing how the feature distribution changes under shifts on $Y$.

**Adaptive agents** are goal-directed systems whose outputs are 'moved by reasons' (Dennett, 1989), meaning they choose an action because they expect it to achieve some desired outcome, and would act differently if they knew that the consequences of their actions would be different. For example, a firm sets prices to maximise profit, and adapts pricing to changes in demand (Kenton et al., 2023).

Consider the transfer learning setting where an agent has to generalise to a target domain using only its previous experience (i.e. zero-shot adaptation), enabled by the fact that it has perfect knowledge of what domain shift has occurred $D_\sigma = \{\sigma\}$ (e.g. the agent conditions their policy on the domain indicator[3] $\pi_\sigma = \pi(d \mid \mathbf{pa}_D, \sigma)$). If the policy $\pi_\sigma$ satisfies a tight regret bound then by Theorem 2 we can reconstruct the underlying CBN from the agent's policy alone (following the procedure described in Appendix C). Hence, any agent that is capable of adapting to known domain shifts has also learned a causal model of their environment.

*Example:* Doctors are one such agent, as they are expected to make low regret decisions under a wide range of known distributional shifts without re-training in the shifted environment. For example, consider a doctor tasked with risk-stratifying patients based on their signs and medical history. The doctor may be transferred to a new ward where patients have received a treatment (known distributional shift) that has a stochastic effect on latent variables (mixed intervention) such as curing diseases and causing side effects. The doctor cannot re-train in this new domain, e.g. randomising decisions and observing outcomes. To be capable of this, Theorem 2 implies the doctor must have learned the causal relations between the relevant latent variables—how the treatment affects diseases, how these diseases and their symptoms are causally related, and so on. Likewise, any medical AI that hopes to replicate this capability must have learned a similarly accurate causal model, and the better the agent's performance the more accurate its causal model must be.

**Causal inference.** Theorem 1 can also be interpreted purely in terms of causal inference. We can compare to the causal hierarchy theorem (CHT) (Bareinboim et al., 2022), which states that an oracle for L1 queries (observational) is almost always insufficient to evaluate all L2 queries (interventional). Our Theorem 1 can be stated in an analogous way; an oracle for optimal policies under mixtures of local interventions $\Pi^*_\Sigma : \sigma \mapsto \pi^*(\sigma)$, can evaluate all L2 queries, which follows from the fact that the oracle identifies the underlying CBN which in turn identifies all L2 queries. Note $\Pi^*_\Sigma$ is a strict subset of L2, and we describe a subset of L2 as being *L2-complete* if evaluating these queries is sufficient to evaluate all L2 queries. Hence Theorem 1 can be summarised as $\Pi^*_\Sigma$ is L2-complete. It would be interesting in future work to determine what other strict subsets of L2 are L2-complete, as identifying these queries is sufficient to identify all interventional queries.

Why is this surprising? Firstly, we may expect the optimal policies to encode a relatively small number of causal relations, as they can be computed from $\mathbb{E}[u \mid d, \mathbf{pa}_D; \sigma]$, which describes the response of a single variable $U$ to intervention $\sigma$. However, Theorem 1 shows that the optimal policies encode all causal and associative relations in $\mathbf{Anc}_U$, including causal relations between latent variables, for example $P(\boldsymbol{Y}_x)$ for any $\boldsymbol{X}, \boldsymbol{Y} \subseteq \mathbf{Anc}_U$. Secondly, Theorems 2 and 3 combined imply that learning to generalise under domain shifts is *equivalent* to learning a causal model of the data generating process—problems that on the surface are conceptually distinct.

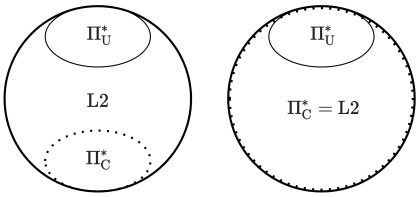



Figure 4: The set L2 (Bareinboim et al., 2022) includes all associative and interventional queries (here restricted to $\mathbf{Anc}_U$), i.e. all queries that can be evaluated given the true CBN over $\mathbf{Anc}_U$. Let $\boldsymbol{\pi}^*_C$ denote the set of queries that can be evaluated with an oracle for optimal policies under all domain shifts for a given $U$. Let $\boldsymbol{\pi}^*_U$ denote the set queries that can be evaluated with an oracle for optimal policies for all task shifts (changing the utility function). It is known that L2 contains $\boldsymbol{\pi}^*_\Sigma$ and $\boldsymbol{\pi}^*_U$ (left, Theorem 3). We show that $\boldsymbol{\pi}^*_C$ almost always contains L2 and hence $\boldsymbol{\pi}^*_C = \text{L2}$ almost always (right, Theorem 1). This implies that $\boldsymbol{\pi}^*_C$ contains $\boldsymbol{\pi}^*_{U'}$ for all $U'$ such that $\mathbf{Pa}_{U'} \in \mathbf{Anc}_U$.

[3]Domain indicator $\sigma$ is equivalent to an environment index (Gupta et al., 2023; Arjovsky et al., 2019)

## 4 DISCUSSION

Here we discuss the consequences for several fields and open questions, as well as limitations.

**Causal representation learning.** Causal representation learning (CRL) aims to learn representations of data that capture unknown causal structure (Schölkopf et al., 2021), with the aim of exploiting causal invariances to achieve better generalisation across domains. Theorems 1 and 2 show that any method that enables generalisation across many domains necessarily involves learning an (approximate) causal model of the data generating process—i.e. a causal representation. Hence, our results provide theoretical justification for CRL by showing it is necessary for strong domain generalisation.

**Causal bounds on transfer learning.** As described in Section 3.2, Theorems 1 and 2 imply fundamental causal constraints on certain transfer learning tasks. For example in the supervised learning task depicted in Figure 1, identifying regret-bounded policies under covariate and label shifts requires learning the causal relations between features and labels. Causal discovery problems such as this are well understood in many settings (Vowels et al., 2022), and in general identifying this causal structure (e.g. that $Y \to X$ in Figure 5 a)) is impossible without interventional data and/or additional assumptions. This connection allows us to convert (im)possibility results for causal discovery to (im)possibility results for transfer learning. Future work could explore this for smaller sets of distributional shifts and derive more general causal bounds on transfer learning.

**Good regulator theorem.** The good regulator theorem is often interpreted as saying that any good controller of a system must have a model of that system (Conant & Ross Ashby, 1970). However, some imagination is needed to take this lesson from the actual theorem, which technically only states that there exists an optimal regulator that is a deterministic function of the state of the system (Wentworth, 2021). Our theorem less ambiguously states that any robust agent must have learned an (approximate) casual model of the environment, as described in Section 3.2. It can therefore be interpreted as a more precise, causal good regulator theorem.

**Emergent capabilities.** Theorems 1 and 2 imply an agent can learn a causal model of the data generating process simply by optimising an objective function $U$ across many domains. By Theorem 3 this causal model could (in principle) enable the agent to identify low-regret policies for any other given objective function $U'$ without additional data, so long as $\mathbf{Pa}_{U'} \in \mathbf{Anc}_U$. This could offer a theoretical explanation for emergent task generalisation, where agents trained on narrow tasks across many domains can generalise to new tasks without retraining (Wang et al., 2020). While this remains speculative, early results suggest that some deep learning models learn causal world models which influence decision making (discussion in Section 5). That causal reasoning could emerge from maximising utility lends support to the hypothesis that simple training objectives can give rise to complex cognitive capabilities (Silver et al., 2021).

**Causal discovery.** Theorems 1 and 2 involve learning the causal structure of the environment by observing the agent's policy under interventions. It is perhaps surprising that the response of this single variable is sufficient to identify all associative and causal relations in $\mathbf{Anc}_U$. In contrast, typical causal discovery algorithms involve measuring the response of many variables to interventions (Vowels et al., 2022). Many causal discovery algorithms make use of the assumption of independent causal mechanisms (Schölkopf et al., 2021), which is equivalent to assuming no agents are present in the data generating process (Kenton et al., 2023). Our results suggest that agents could be powerful resources for causal discovery. In Appendix B we use our results to construct a causal discovery algorithm for learning the causal structure over latents, and test it on synthetic data.

**Applicability of causal methods.** Causal models have been used to formally define concepts such as intent (Halpern & Kleiman-Weiner, 2018), harm (Richens et al., 2022), deception (Ward et al., 2023b), manipulation (Ward et al., 2023a) and incentives (Everitt et al., 2021), and are required for approaches to explainability (Wachter et al., 2017) and fairness (Kusner et al., 2017). These definitions, and methods for training safe and ethical AI systems require causal models of the data generating process, which are typically hard to learn, leading some to doubt their practicality (Fawkes et al., 2022; Rahmattalabi & Xiang, 2022). However, our results show that by training increasingly robust and capable agents we also learn increasingly accurate causal models, which (as we demonstrate) can be elicited from those agents and could be used to support these causal methods.

**Limitations.** The most significant limitation of Theorems 1 and 2 is that they require agents to be robust to a large set of domain shifts (i.e. mixtures over local interventions). Theorem 2 implies the effect of loosening regret bounds for certain shifts is that some causal relations become unidentifiable

from the agent's policy. Hence, we expect that for agents that are robust to a smaller set of domain shifts we will be able to identify a less complete causal model of the environment. Another limitation is that our results only apply to unmediated decision tasks (Assumption 1). We expect Theorems 1 and 2 can be extended to active decision tasks, as Assumption 1 does not play a major role beyond simplifying the proofs. Finally, to derive our results it was necessary to restrict to environments with finite-dimensional variables.

## 5 Related work

Recent empirical works have explored the question of if deep learning models learn 'surface statistics' (e.g. correlations between inputs and outputs) or learn internal representations of the world (McGrath et al., 2022; Abdou et al., 2021; Li et al., 2022; Gurnee & Tegmark, 2023). Our results offer some theoretical clarity to this discussion, tying an agents performance to the accuracy its world model, and showing that agents that learn only 'surface statistics' have limited robustness. One study in particular (Li et al., 2022) found that a GPT model trained to predict legal next moves in the board game Othello learned a linear representation of the board state (Nanda, 2023). Further, by intervening on the intermediate activations they could change the internally represented board state, with the model updating its predictions consistently with these interventions, including when the post-intervention board state lay outside of the training distribution. This indicates that the model is learning and utilising a representation of the data generating process that can support out-of-distribution generalisation under interventions—much like a causal model.

The problem of evaluating policies under distributional shifts has been studied extensively in causal transportability (CT) theory (Bareinboim & Pearl, 2016; Bellot & Bareinboim, 2022). CT aims to provide necessary and sufficient conditions for policy evaluation under distributional shifts when all assumptions on the data generating process (i.e. inductive biases) can be expressed as constraints on causal structure (Bareinboim & Pearl, 2016). However, deep learning algorithms can exploit a much larger set of inductive biases (Neyshabur et al., 2014; Battaglia et al., 2018; Rahaman et al., 2019; Goyal & Bengio, 2022) which in many real-world tasks may be sufficient to identify low regret policies without requiring causal knowledge. Thus, CT does not imply that agents must learn causal models in order to generalise unless we assume agents only use causal assumptions to begin with, which amounts to proof by assumption. See Appendix G for further discussion.

A similar result to Theorems 1 and 2 is the causal hierarchy theorem (CHT) (Bareinboim et al., 2022; Ibeling & Icard, 2021), which shows that observational data is almost always *insufficient* for identifying all causal relations between environment variables, whereas our results state that the set of optimal policies is almost always sufficient to identify all causal relations. In Section 3.2 we discuss the similarities between these theorems, and in Appendix G we discuss the differences.

## 6 Conclusion

Causal reasoning is foundational to human intelligence, and has been conjectured to be necessary for achieving human level AI (Pearl, 2019). In recent years, this conjecture has been challenged by the development of artificial agents that achieve strong generalisation without explicitly learning or reasoning on causal models. And while the necessity of causal models for solving causal inference tasks has been established (Bareinboim et al., 2022), their role in decision tasks such as classification or reinforcement learning has been less clear.

We have resolved this conjecture in a model-independent way, showing that any agent capable of robustly solving a decision task must have learned a causal model of the data generating process, regardless of how the agent is trained or the details of its architecture. This hints at an even deeper link between causality and general intelligence, as the causal model the agent learns can in turn support a much wider range of tasks beyond the original training objective. Conversely, it shows that causal models can be learned simply by maximising utility (or minimising loss) across many environments. By establishing this formal connection between causality and generalisation, our results show that causal reasoning is a fundamental ingredient for developing robust and general AI.

**Acknowledgements.** We would like to thank Alexis Bellot, Damiano Fornasiere and Pietro Greiner for their helpful comments on the manuscript.

## REFERENCES


Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. *arXiv preprint arXiv:2109.06129*, 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pp. 100–108. PMLR, 2012a.

Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 698–704, 2012b.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. 2022.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Alexis Bellot and Elias Bareinboim. Partial transportability for domain generalization. 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Roger C Conant and W Ross Ashby. Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97, 1970.

Juan Correa and Elias Bareinboim. 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, 2020.

A P Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review / Revue Internationale de Statistique*, 70:161–189, 2002.

Daniel C Dennett. *The intentional stance*. MIT press, 1989.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 12004–12019. PMLR, 2022.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pp. 1802–1811. PMLR, 2019.

Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11487–11495, 2021.

Benjamin Eysenbach and Sergey Levine. Maximum entropy RL (provably) solves some robust RL problems. *arXiv preprint arXiv:2103.06257*, 2021.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

Jake Fawkes, Robin Evans, and Dino Sejdinovic. Selection, ignorability and challenges with causal fairness. In *Conference on Causal Learning and Reasoning*, pp. 275–289. PMLR, 2022.

Alison Gopnik, Laura Schulz, and Laura Elizabeth Schulz. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press, 2007.

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022.

Sharut Gupta, Stefanie Jegelka, David Lopez-Paz, and Kartik Ahuja. Context is environment, 2023.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Joseph Halpern and Max Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Ronald A Howard and James E Matheson. Influence diagrams. *Decision Analysis*, 2(3):127–143, 2005.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Duligur Ibeling and Thomas Icard. A topological perspective on causal inference. *Advances in Neural Information Processing Systems*, 34:5608–5619, 2021.

Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering agents. *Artificial Intelligence*, pp. 103963, 2023.

Daphne Koller and Brian Milch. Multi-agent influence diagrams for representing and solving games. *Games and economic behavior*, 45(1):181–221, 2003.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17:391–444, 2007.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in AlphaZero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119, 2022.

Christopher Meek. Strong completeness and faithfulness in bayesian networks. *arXiv preprint arXiv:1302.4973*, 2013.

Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pp. 6–10. IEEE, 2018.

Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Causal inference via kernel deviance measures. *Advances in neural information processing systems*, 31, 2018.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Neel Nanda. Actually, othello-gpt has a linear emergent world model, Mar 2023. URL <https://neelnanda.io/mechanistic-interpretability/othello>.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Masashi Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, pp. 763–765, 1973.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition edition, 2009. ISBN 9780521895606.

Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 247–254, 2011.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810. IEEE, 2018.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

Aida Rahmattalabi and Alice Xiang. Promises and challenges of causality for ethical machine learning. *arXiv preprint arXiv:2201.10683*, 2022.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1956.

Jonathan Richens, Rory Beard, and Daniel H Thompson. Counterfactual harm. *Advances in Neural Information Processing Systems*, 35:36350–36365, 2022.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. Causally regularized learning with agnostic data selection bias. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 411–419, 2018.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

Steven A Sloman and David Lagnado. Causality in thought. *Annual review of psychology*, 66: 223–247, 2015.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Francis Rhys Ward, Tom Everitt, Francesco Belardinelli, and Francesca Toni. Honesty is the best policy: defining and mitigating AI deception. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli. Defining deception in structural causal games. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 2902–2904, 2023b.

John Wentworth. Fixing the good regulator theorem. https://www.alignmentforum.org/posts/Dx9LoqsEh3qHNJMDk/fixing-the-good-regulator-theorem, 2021. Accessed: 2023-10-17.

Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.