Meta-reinforcement learning of causal strategies

Ishita Dasgupta[*1,4], **Zeb Kurth-Nelson**[1,2]
Silvia Chiappa[1], **Jovana Mitrovic**[1], **Pedro Ortega**[1],
Edward Hughes[1], **Matthew Botvinick**[1,3], **Jane Wang**[1]

[1]DeepMind, UK
[2]MPS-UCL Centre for Computational Psychiatry, UCL, UK
[3]Gatsby Computational Neuroscience Unit, UCL, UK
[4]Department of Physics and Center for Brain Science, Harvard University, USA

Abstract

Discovering and exploiting the causal structure in the environment is a crucial challenge for intelligent agents. However, it is not understood how such reasoning comes about, even in natural intelligence. Here, we investigate the emergence of causal reasoning and intervention strategies from simpler reinforcement learning algorithms using a meta-reinforcement learning framework. We find that, after training on distributions of environments having causal structure, meta-learning agents learn to perform a form of causal reasoning in related, held-out tasks. In particular, we find that the form of causal reasoning learned relates to the information encountered during learning, ranging from causal inference from observations, to resolving confounders, selecting informative interventions, and making counterfactual predictions. Empirical findings in human behavioral research suggest promising connections between our model and the development and implementation of causal reasoning in humans.

1 Introduction

Empirical work in human developmental research suggests that humans' ability to perform causal reasoning emerges through experiences in the world rather than from an innate theory of causality [3, 4, 21, 30]. The question then arises of what learning mechanisms allow humans to acquire this ability through experience. In this work, we demonstrate how causal reasoning can arise in agents via meta-learning over tasks that contain causal structure. In particular, we use a "meta-reinforcement learning" framework [6, 34] that enables performing *interventions* in the environment, an essential ingredient for causal reasoning. This methodology has previously been shown to give rise to complex policies that exploit structure in the task distribution [23, 29, 34, 35].

A hallmark of learning to reason about cause and effect from experience is that the (causal) inference algorithm learned should reflect the structure of the environment. If normative causal reasoning provides an advantage, and is possible given the data and the structure of the environment, then an agent should be able to learn it. However, certain other experiences might lead to different algorithms that vary on the spectrum of how "causally-aware" they are. Graded-ness of causal inference is observed in adult humans, with humans showing characteristic deviations from normative inference, often tending toward associative reasoning [7, 8, 27, 28], with causal notions varying significantly with domain and function [17, 19]. Learning causality from experience, as in our framework, offers a possible explanation – different experiences potentially support different kinds and extents of causal reasoning. In this paper, we test these predictions in 5 experiments. We see that architecturally identical agents can learn different strategies for reasoning about causal structure depending on the kinds of experiences gathered during training.

[*]Corresponding author: ishitadasgupta@g.harvard.edu

3rd Workshop on Meta-learning at NeurIPS 2019, Vancouver, Canada.

Finally, formal approaches to causal identification (determining the causal graph from data) often require large amounts of data [9, 31, 33], and inference in the constructed causal graphs is also computationally expensive [16]. In real-world environments, humans operate under time, data, and resource constraints, dealing with uncertainty in model structure as well as non-stationarity. Agents that learn aspects of the learning algorithm directly from experience will adapt to statistical structure in their specific environment and task, and could utilize useful abstract priors (or inductive biases) from other episodes that can be difficult to formally specify. Such adaptations amortize much of the computation over previous experience and could allow better performance than formal approaches under ecological constraints [e.g. 5, 10, 11, 18, 32].

The purpose of this work is not to propose a new algorithmic solution to causal inference per se. Rather, we argue that our meta-learning approach has compelling links to human causal reasoning in terms of a) how a theory of causality could be learned, b) the graded notion of causality in humans, and c) resource efficiency by meta-learning inductive biases. Such resource efficient causal inference is also useful for machine learning [e.g. 1, 13, 20, 22, 24] (see also supplementary material).

2 Task Setup

In our experiments, we use a simple framework that shares key properties with human causal reasoning. First, the number of variables over which inference is carried out is small. Second, the amount of data available (in each task) is limited. Third, agents can actively seek out information by interacting with the environment rather than only receiving passive input. In each episode the agent interacts with a different Causal Bayesian Network (CBN) $\mathcal{G}$ with $N = 5$ variables. The structure of $\mathcal{G}$ is drawn randomly with some constraints. Each episode consists of $T = 5$ steps (i.e. very limited interaction within each task), which are divided into two phases: an *information phase* and a *quiz phase*.

The information phase corresponds to the first $T-1$ steps during which the agent performs information-gathering actions and sees node values sampled from $\mathcal{G}$ (one node is hidden to allow for unobserved confounders). Note that $\mathcal{G}$ is never directly provided to the agent, but is only observed through $T-1$ samples. An example information phase action is choosing a node to intervene upon, and observing the resulting values of other nodes. Agents have to perform two distinct tasks during the information phase: a) choose information gathering actions, and b) process the resulting data to allow downstream causal reasoning (in quiz phase, see below). To better understand (a), we include a *random* condition (as opposed to an *active* condition) where the environment ignores the agents information phase actions and randomly chooses information gathering actions. To better understand (b), different agents have access to different kinds of data for the same information phase action (as detailed in the experiments below). All the agents in the different experiments are architecturally identical, and give rise to different behavior solely due to differences in the data they receive in the information phase. The quiz phase, corresponding to the final step T, requires the agent to exploit the causal knowledge accumulated during the information phase, to select the node with the highest value under an external intervention where a randomly selected node is set to -5. Since the quiz phase requires the agent to predict the outcome of a previously unseen intervention, consistently good performance in general requires causal reasoning. The structure of the quiz phase is exactly the same for all agents in all experiments.

We train a recurrent agent (via A3C; see supplementary material for details) on many different such CBNs. We test on held out CBNs that have not been seen before with learning turned off. The agent has to implement a causally-aware strategy to learn about the new CBN in an episode and perform well in the quiz phase. This strategy itself is meta-learned across many previously encountered CBNs.

3 Experiments

Our two experiments differ in the kinds of data the agent receives in response to the same information phase actions (detailed below), although the quiz phase is the same. Agent performance is measured as the reward earned in the quiz phase for held-out CBNs, normalized by the maximum possible reward achievable with exact causal reasoning. Choosing a random node in the quiz phase gives an average reward of $-5/4 = -1.25$ since the externally intervened node always has value -5 and the others have on average 0 value. We train 8 copies of each agent and report the average performance across 1632 test episodes. Error bars indicate 95% confidence intervals.

Observational Response: In the information phase, the actions of the agent are ignored. The agent always receives the values of the visible nodes sampled from the joint distribution of $\mathcal{G}$. As the default

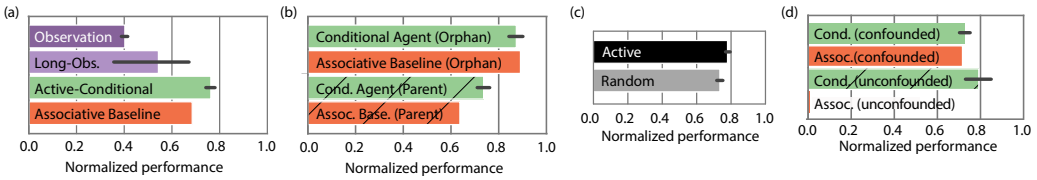

Figure 1: **Experiment 1. Agents exhibit causal strategies from observational data.** a) Average normalized performance of the agents. b) Performance split by the presence or absence of at least one parent (Parent and Orphan respectively) on the externally intervened node. c) Performance for Active vs Random Conditional agents. d) Performance for Associative Baseline vs Active Conditional agents, where intervened node has a parent.

$T=5$ episode length can prove quite challenging, we also train an agent with $4\times$ longer episode length (Long Observational) to measure performance with access to more data.

Conditional Response: This provides more informative observations. Specifically, the agent's information phase actions correspond to observing a world in which the selected node X_j is set to a high value (5) outside the likely range of normal observations, and the remaining nodes are sampled from the conditional distribution $p(X_{1:N\setminus j}|X_j=5)$. We run active and random versions.

Interventional Response: The information phase actions correspond to setting the selected node X_j to a high value (5) and sampling the remaining nodes from the *intervened distribution* $p_{\rightarrow X_j=5}(X_{1:N\setminus i}|X_j=5)$ (see Supplementary material). We run active and random versions.

An intuitive example for how these environmental responses differ is the following: when gauging whether smoking causes cancer, the different kinds of responses differ in the data they provide. The observational response provides a uniformly random set of people, and their smoking behavior, and cancer rates; the conditional response selectively provides "informative samples" in the form of people who either never smoke or smoke a lot along with corresponding cancer rates; and the interventional response reports the results of a randomized controlled trial in which people were assigned at random to smoke or not. Note that the causal inferences in our task are over more nodes, and the conditional and interventional responses also require specification of the node on which to condition or intervene. Various kinds of causal inferences can be made from these different kinds of data [25]. Significant causal reasoning beyond correlations is possible from passive observations alone (studied in Experiment 1). However, in the presence of hidden confounders, interventions might be required (studied in Experiment 2).

3.1 Experiment 1: Observational Environments

In Experiment 1, the agents are in environments that do not permit any interventions during the information phase (only observational and conditional responses). We show that, even under these restrictions, agents are able to demonstrate behavior consistent with learning about cause and effect to some extent. To demonstrate this, we compare agents' performance with that of an "Associative Baseline". This baseline acts solely on correlational information, i.e. it chooses the node that has the maximum value as per the exact $p(X_j|X_i=-5)$, with X_i the node externally intervened upon. We also study the *kind* of causal inference learned, and the role of actively choosing information phase actions.

Results: We find that when given access to informative observations, our agents learn to perform causal reasoning from observational data – the agent receiving Active-Conditional responses significantly outperforms the Associative Baseline (Fig. 1a). To further demonstrate that this improvement is due to causal reasoning, we partition the test cases by whether or not the node that was intervened on in the quiz phase has a parent (Fig. 1b). If the intervened node X_j has no parents, then $\mathcal{G}$ is the same as the CBN in which X_j has been intervened upon, $\mathcal{G}_{\rightarrow X_j}$, and causal reasoning affords no advantage over associative reasoning. Indeed, the Active-Conditional agent performs better than the Associative Baseline only when the intervened node has parents (Fig. 1b). Agents that receive unconditional observations, i.e. Observational responses (Fig. 1a) perform worse than with Active-Conditional responses, as expected since these provide less diagnostic information. However, they still perform better than the random action baseline and the same agent learns to utilize more data (Long-Obs.) to yield better performance. We find that the Active-Conditional agent's performance is slightly but significantly ($p=0.003$, Fig. 1c) higher than the Random-Conditional agent. This indicates that when permitted, the agent learns to select actions that provide informative observations.

The Active-Conditional agent however does not utilize full causal reasoning (= 1.0 performance on our scale). From Fig. 1b, we see that this drop is driven mostly by test cases where the intervened node has a

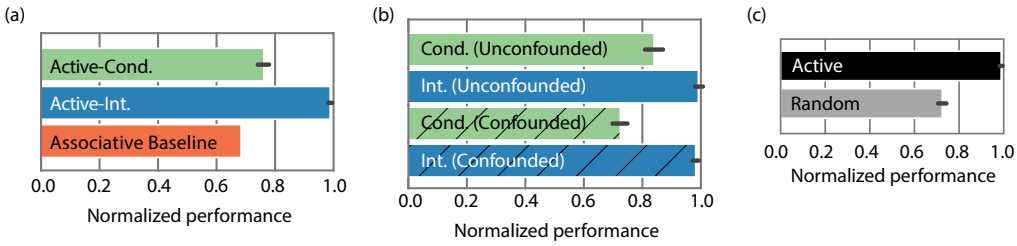

Figure 2: **Experiment 2. Agents exhibit causal strategies from interventional data.** a) Average normalized performance of the agents. b) Performance split by the presence or absence of unobserved confounders (abbreviated as Conf. and Unconf.). c) Performance of active vs passive interventional agents.

parent. We hypothesize that this is due to the presence of unobserved confounders. We consider the test cases where the externally intervened node has parents and partition into confounded parent(s) or unconfounded parent(s) (Fig. 1d). We see that the performance of the Active-Conditional agent is significantly higher than the associative baseline only in cases where the parent is unconfounded. In confounded cases, it is not in general possible to do causal inference with access to only observational data. In the next experiment, we discuss the performance of our agents in an environment that permits interventions.

3.2 Experiment 2: Interventional Environments

In this experiment, we test if agents can learn to perform causal inference from interventions by allowing interventional responses to information phase actions. In particular, we are interested in performance in the presence of unobserved confounders. We test both active and random versions of the agent.

Results: We see in Fig. 2a that the agent with access to Active-Interventional responses from the environment performs better than the Active-Conditional agent, achieving close to optimal performance. This shows that when given access to interventions, the agent learns to leverage them to perform causal reasoning. Partitioning the test cases by whether the externally intervened node has unobserved confounders (Fig. 2b), we see that the Active-Interventional agent performs close to optimal on both confounded and unconfounded test cases. Further, we find that the Active-Interventional agent learns to strategically control the interventions performed and choose highly informative interventions: its performance is significantly better than the Random-Interventional agent and almost at ceiling (Fig. 2c).

Even restricted to inference in the absence of confounders however, the performance of the Active-Conditional agent is not as high as the performance of the Interventional Agent – even though such inferences are in theory within reach of the conditional agent. This could be an example of the agent utilizing unspecified statistical information – in our framework, the final quiz phase node values are the negative (with noise) of the values observed, if the quiz phase node is intervened on in the information phase. Given interventional responses observed in information phase therefore, it is relatively easy to predict the effects of the random external intervention in quiz phase (see supplementary material for more complex cases). But with access to only conditional observations, one has to remember and integrate information across several values observed in information phase to correctly predict the result of the external intervention in the quiz phase. When utilizing such structure, interventions are easier to learn from as also observed in humans [8, 7].

4 Discussion

Mirroring the criteria for valuable links to human causal reasoning outlined in the introduction, we show that a) causal reasoning capabilities can be learned via meta-learning through interaction with an environment that rewards and permits causal reasoning, and b) graded kinds and extents of causal reasoning can arise depending on the data we have access to. We find that depending on the environment, our agents learn to: 1) leverage observational data to make causal inferences, 2) leverage interventions to resolve unobserved confounders, and 3) actively generate informative data. In the supplementary, we showcase our agents performing counterfactual reasoning. And finally, c) even in this simple domain, we observe evidence of unspecified, non-trivial underlying statistical structure in the environment, as well as preliminary evidence that our agents utilize it to amortize and simplify inferences. Other work on statistical approaches to learning causal structure [1, 15, 14], as well as methods from neuroscience [35], could provide further insights into what our agents learn.

A crucial contribution of our work is to consider causal reasoning in natural intelligence not an end in and of itself but a means to better performance on some downstream task that is easier to specify, in

a world that contains causal structure. In our case this task is acquiring reward in an RL task, but could be generalized to any other task by simply changing the meta-learning objective. This is a reasonable assumption since causal reasoning exists in humans, and even chimpanzees and rats [2, 12, 26] without "formal instruction" on causality itself. This assumption allows us to frame the acquisition of causal reasoning as a meta-learning problem, and we highlight how this approach could also capture many qualitative empirical findings in how causal reasoning is learned and implemented in humans.

References

[1] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

[2] A. Blaisdell, K. Sawa, K. Leising, and M. Waldmann. Causal reasoning in rats. *Science*, 311(5763):1020–1022, 2006.

[3] E. B. Bonawitz, D. Ferranti, R. Saxe, A. Gopnik, A. N. Meltzoff, J. Woodward, and L. E. Schulz. Just do it? investigating the gap between prediction and action in toddlers' causal inferences. *Cognition*, 115(1):104–117, 2010.

[4] S. Carey. *The origin of concepts*. Oxford University Press, 2009.

[5] I. Dasgupta, E. Schulz, J. B. Tenenbaum, and S. J. Gershman. A theory of learning to infer. *bioRxiv*, 2019.

[6] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

[7] P. M. Fernbach, A. Darlow, and S. A. Sloman. Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, 21(3):329–336, 2010.

[8] P. M. Fernbach and B. Rehder. Cognitive shortcuts in causal inference. *Argument & Computation*, 4(1):64–88, 2013.

[9] D. Geiger, T. Verma, and J. Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990.

[10] S. J. Gershman, E. J. Horvitz, and J. B. Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015.

[11] G. Gigerenzer and H. Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, 1(1):107–143, 2009.

[12] A. Gopnik, C. Glymour, D. Sobel, L. Schulz, T. Kushnir, and D. Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.

[13] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[14] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

[15] D. Janzing, P. O. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. *arXiv preprint arXiv:0909.4386*, 2009.

[16] M. I. Jordan and Y. Weiss. Graphical models: Probabilistic inference. *The handbook of brain theory and neural networks*, pages 490–496, 2002.

[17] T. R. Krynski and J. B. Tenenbaum. The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3):430, 2007.

[18] F. Lieder and T. L. Griffiths. Strategy selection as rational metareasoning. *Psychological Review*, 124:762–794, 2017.

[19] T. Lombrozo. Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61(4):303–332, 2010.

[20] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.

[21] A. N. Meltzoff. Infants' causal learning: Intervention, observation, imitation. 2007.

[22] J. Mitrovic, D. Sejdinovic, and Y. W. Teh. Causal inference via kernel deviance measures. In *Advances in Neural Information Processing Systems*, pages 6986–6994, 2018.

[23] P. A. Ortega, J. X. Wang, M. Rowland, T. Genewein, Z. Kurth-Nelson, R. Pascanu, N. Heess, J. Veness, A. Pritzel, P. Sprechmann, et al. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.

[24] G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf. Learning independent causal mechanisms. *arXiv preprint arXiv:1712.00961*, 2017.

[25] J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

[26] D. Premack and A. J. Premack. Levels of causal understanding in chimpanzees and children. *Cognition*, 50(1-3):347–362, 1994.

[27] B. Rehder. Independence and dependence in human causal reasoning. *Cognitive psychology*, 72:54–107, 2014.

[28] B. Rehder and M. R. Waldmann. Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & cognition*, 45(2):245–260, 2017.

[29] S. Ritter, J. X. Wang, Z. Kurth-Nelson, S. M. Jayakumar, C. Blundell, R. Pascanu, and M. Botvinick. Been there, done that: Meta-learning with episodic recall. *arXiv preprint arXiv:1805.09692*, 2018.

[30] R. Saxe and S. Carey. The perception of causality in infancy. *Acta psychologica*, 123(1-2):144–165, 2006.

[31] P. Spirtes, C. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000.

[32] P. M. Todd and G. Gigerenzer. Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, 16(3):167–171, 2007.

[33] T. Verma and J. Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

[34] J. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *CoRR*, abs/1611.05763, 2016.

[35] J. X. Wang, Z. Kurth-Nelson, D. Kumaran, D. Tirumala, H. Soyer, J. Z. Leibo, D. Hassabis, and M. Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860, 2018.

Supplementary to Meta-reinforcement learning of causal strategies

1 Related Work

Goodman et al. [13] demonstrated how an abstract notion of causality in humans can be learned from experience, with hierarchical Bayesian inference. Our approach is similar to this as meta-learning can also be framed as hierarchical Bayesian inference [14]. However, these approaches provide complementary advantages. While formal theory learning (as in [13]) is systematic and generalizes across domains, it requires the pre-specification of discrete primitives and an expensive zero order (stochastic search) optimization to learn the correct theory built from these primitives [5, 28]. A restrictive choice of primitives limits the space of possible theories, while a generous choice makes the optimization very expensive. This approach also leaves open the question of the origin of these discrete primitives and how they might be plausibly implemented in the brain. Our method avoids these assumptions and instead uses a first order (gradient-based) optimization method that leverages learning signals from the environment, thus discovering emergent structure directly from experience [19]. Since our model is implemented with a deep neural network, which can be universal approximators [17, 29], it can implement different graded causal theories that don't conform to purely normative accounts, in a neurally-plausible distributed representation. This could give rise to graded causal reasoning behaviors analogous to those seen in humans [10, 11, 25, 26] .

Bengio et al [3] propose a meta-learning approach to utilize explicit, pre-specified statistical properties of interventions to isolate and disentangle causal variables in a supervised learning setting. Our work shows how a spectrum of 'causally-aware algorithms' can arise from utilizing several different kinds of implicit, unspecified statistical structure in the environment. Our reinforcement learning approach further allows the agent to directly interact with the environment to also simultaneously learn an experimental policy that utilizes this underlying structure. Denil et al [8] showed that deep reinforcement learning agents can learn to perform actions to gain knowledge about latent, physical properties of objects, but do not explore explicit causal inference.

2 Problem Specification

Our goal is to demonstrate that causal reasoning can arise from meta-reinforcement learning. Further, we demonstrate that depending on the kinds of data the agents see during training, the kind of causal reasoning learned varies. Our agents learn to leverage statistical structure in different kinds of available information, to carry out different kinds of causal reasoning. In this section, we first briefly formalize causal inference and how it depends on the kinds of data the environment provides.

Causal relationships among random variables can be expressed using *causal Bayesian networks* (CBNs $\mathcal{G}$) [7, 23, 30]. Each node X_i corresponds to a random variable, and the joint distribution $p(X_1,...,X_N)$ is given by the product of conditional distributions of each node X_i given its parent nodes $\text{pa}(X_i)$, i.e. $p(X_{1:N}) = \prod_{i=1}^{N} p(X_i|\text{pa}(X_i))$.

The edges of $\mathcal{G}$ encode causal semantics: a directed path from X_c (cause) to X_e (effect) is called a causal path. The causal effect of X_c on X_e is the conditional distribution of X_c given X_e restricted to only causal paths. This restriction is an essential caveat, since the simple conditional distribution $p(X_e|X_c)$ encodes only correlations (i.e. associative reasoning). Intervening on a node X_c corresponds

3rd Workshop on Meta-learning at NeurIPS 2019, Vancouver, Canada.

to removing its connection to its parent nodes pa(X_c), and fixing it to some value C yielding a new CBN $\mathcal{G}_{\rightarrow X_c=C}$. The causal effect of X_c on X_e is given by the conditional distribution in this new CBN. This distribution is denoted as $p_{\rightarrow X_c=C}(X_e|X_c=C)$.

Different kinds of environments support different kinds of causal reasoning. It is often possible to compute $p_{\rightarrow X_c=C}(X_e|X_c=C)$ (i.e. causal reasoning) using observations from $\mathcal{G}$[1]. We investigate this kind of causal reasoning in Experiment 1 (Observational Environments). However, in the presence of unobserved confounders (an unobserved variable that affects both X_c and X_e), this is, in general , no longer possible [23]. The only way to compute causal effects $p_{\rightarrow X_c=C}(X_e|X_c=C)$ in this case is by collecting observations directly from the intervened graph $\mathcal{G}_{\rightarrow X_c=C}$. In Experiment 2 (Interventional Environments), we investigate this kind of causal reasoning, by allowing agent to perform interventions on the environment. An additional level of sophistication comes from *counterfactual* environments (see section 7.1 for results).

3 Extended Methods

Causal Graphs, Observations, and Actions

We generate graphs that have $N=5$ nodes and sample the adjacency matrix to have non-zero entries only in its upper triangular part (this guarantees that all the graphs obtained are acyclic). Edge weights w_{ji} are uniformly sampled from $\{-1,0,1\}$. This yields $3^{N(N-1)/2}=59049$ unique graphs. These can be divided into equivalence classes, i.e. sets of graphs that are structurally identical but differ in the permutation order of the node labels. Our held-out test set consists of 12 random graphs plus all other graphs in the corresponding equivalence classes, yielding 408 total graphs in the test set. Each of these are tested with 4 possible external interventions giving a total of 1632 test episode. Thus, none of the graphs in the test set (or any graphs equivalent to these) have been seen during training.

We sample each node, $X_i \in \mathbb{R}$, as a Gaussian random variable. The distribution of parentless nodes is $\mathcal{N}(\mu=0.0,\sigma=0.1)$, while for a node X_i with parents pa(X_i) we use the conditional distribution $p(X_i|\text{pa}(X_i))=\mathcal{N}(\mu=\sum_j w_{ji}X_j,\sigma=0.1)$ with $X_j \in \text{pa}(X_i)$. We also tested graphs with non-linear causal effects and larger graphs of size N = 6 (see Section 7.3).

A root node of $\mathcal{G}$ is always hidden, to allow for unobserved confounders, and the agent can therefore only ever see the values of the other 4 nodes. These 4 nodes are henceforth referred to as the 'visible nodes'. The concatenated values of the nodes, v_t, and a one-hot vector indicating the external intervention during the quiz phase, m_t, (explained below) form the observation vector provided to the agent at step t, $o_t=[v_t,m_t]$[2].

In both phases, at each step t, the agent chooses to take one out of $2(N-1)$ actions. The first $N-1$ actions are *information actions*, and the second $N-1$ actions are *quiz actions*. Both information and quiz actions are associated with selecting the $N-1$ visible nodes, but can only be legally used in the appropriate phase of the task. If used in the wrong phase, a penalty is applied and the action produces no effect.

Information Phase. The information phase differs depending on the kind of environment the agent is in – observational or interventional. Here, we discuss the case of the interventional environment.

An information action $a_t=i$ causes an intervention on the i-th node, setting the value of $X_{a_t}=X_i=5$ (the value 5 is outside the likely range of sampled observations and thus facilitates learning the causal graph). The node values v_t are then obtained by sampling from $p_{\rightarrow X_i=5}(X_{1:N\setminus i}|X_i=5)$ (where $X_{1:N\setminus i}$ indicates the set of all nodes except X_i), i.e. from the intervened CBN $\mathcal{G}_{\rightarrow X_{a_t}=5}$. If a quiz action is chosen during the information phase, it is ignored, i.e. the node values are sampled from $\mathcal{G}$ as if no intervention has been made. Furthermore, the agent is given a penalty of $r_t=-10$ in order to encourage it to take quiz actions during the quiz phase. There is no other reward during the information phase.

The default length an episode is fixed to be $T=N=5$, giving an information phase of length of $T-1=4$. This episode length was chosen because in the noise-free limit, a minimum of $N-1=4$ interventions, one on each visible node, is required in general to resolve the causal structure.

[1] When the CBN $\mathcal{G}$ is known, this process can be formalized as do-calculus [23, 24]. In our case the CBN is not directly provided, and the agent must simultaneously perform causal identification using samples from $\mathcal{G}$ [15].

[2] 'Observation' o_t refers to the reinforcement learning term, i.e. the input from the environment to the agent. This is distinct from observations in the causal sense which we refer to as observational data.

Quiz Phase. The quiz phase remains the same for all the different environments and agents. In the quiz phase, one visible node X_j is selected at random to be intervened on by the environment. Its value is set to -5. We chose -5 to disallow the agent from memorizing the results of interventions in the information phase (which are fixed to $+5$) in order to perform well on the quiz phase. The agent is informed which node received this external intervention via the one-hot vector m_t as part of the observation from the the final pre-quiz phase timestep, $T-1$. For steps $t < T-1$, m_t is the zero vector. The agent's reward on this step is the sampled value of the node it selected during the quiz phase. In other words, $r_T = X_i = X_{a_T-(N-1)}$ if the action selected is a quiz action (otherwise, the agent is given a penalty of $r_T = -10$).

Active vs Random Conditions. Our agents have to perform two distinct tasks during the information phase: a) actively choose which nodes to act on and b) perform casual reasoning based on the observations. We refer to this setup as the "active" condition. To better understand the role of (a), we include comparisons with a baseline agent in the "random" condition where the environment ignores the agents actions and randomly chooses a visible node to intervene upon at each step of the information phase. Note again that the only difference between agents in these two conditions is the kind of data the environment provides them.

Two Kinds of Learning. An "inner loop" of learning occurs within each episode where the agent is learning from the 4 samples it gathers during the information phase to perform well in the quiz phase. The same agent then enters a new episode, where it has to repeat the task on a different CBN. Test performance is reported on CBNs that the agent has never previously seen after all the weights of the RNN have been fixed. Hence, the only transfer from the training to test set (or the "outer loop" of learning) is a learned procedure for collecting evidence in the information phase to perform well in the quiz phase. Exactly what this learned procedure is will depend on the training environment. We will show that this learned procedure can include performing different kinds of causal inference, as well as active information gathering.

Agent Architecture and Training.

We used a long short-term memory (LSTM) network [16] (with 192 hidden units) that, at each time-step t, receives a concatenated vector containing $[o_t, a_{t-1}, r_{t-1}, m_t]$ as input, where o_t is the observation, a_{t-1} is the previous action, r_{t-1} the previous reward and m_t indicates the external intervention. The outputs, calculated as linear projections of the LSTM's hidden state, are a set of policy logits (with dimensionality equal to the number of available actions), plus a scalar baseline. The policy logits are transformed by a softmax function, and then sampled to give a selected action.

Learning was by *asynchronous advantage actor-critic* [20]. In this framework, the loss function consists of three terms – the policy gradient, the baseline cost and an entropy cost. The baseline cost was weighted by 0.05 relative to the policy gradient cost. The weighting of the entropy cost was annealed over the course of training from 0.25 to 0. Optimization was via RMSProp with $\epsilon = 10^{-5}$, momentum = 0.9 and decay = 0.95. Learning rate was annealed from 9×10^{-6} to 0, with a discount of 0.93. Hyperparameters were optimized by performing a coarse grid search (2-4 values) over learning rate, discount factor, and the number of hidden units in the LSTM. Unless otherwise stated, training was done for 1×10^7 steps using batched environments with a batch size of 1024, using a distributed architecture.

In all experiments, the agent is tested with the learning rate set to zero using a held-out test set as discussed in the main text.

4 Formalism for Memory-based Meta-learning

Meta-learning refers to a broad range of approaches in which aspects of the learning algorithm itself are learned from the data. Many individual components of deep learning algorithms have been successfully meta-learned, including the optimizer [1], initial weight parameters, [12], a metric space [31], and use of external memory [27].

Following the approach of [9, 32], the entire inner loop of learning is implemented by a recurrent neural network (RNN), and we train the weights of the RNN with model-free reinforcement learning (RL). The RNN is trained on a broad distribution of problems which each require learning. Consider a distribution $\mathcal{D}$ over Markov Decision Processes (MDPs). We train an agent with memory (in our case an RNN-based agent) on this distribution. In each episode, we sample a task $m \sim \mathcal{D}$. At each step t within an episode, the agent sees an observation o_t, executes an action a_t, and receives a reward r_t. Both a_{t-1} and r_{t-1} are given as additional inputs to the network. Thus, via the recurrence of the

network, each action is a function of the entire trajectory $\mathcal{H}_t = \{o_0, a_0, r_0, ..., o_{t-1}, a_{t-1}, r_{t-1}, o_t\}$ of the episode. Because this function is implemented by the neural network, its complexity is limited only by the size of the network. When trained in this way, the RNN is able to implement a learning algorithm capable of efficiently solving novel learning problems in or near the training distribution.

Learning the weights of the RNN by model-free RL can be thought of as the "outer loop" of learning. The outer loop shapes the weights of the RNN into an "inner loop" learning algorithm, which plays out in the activation dynamics of the RNN and can continue learning even when the weights of the network are frozen. The inner loop algorithm can also have very different properties from the outer loop algorithm used to train it. For example, this approach has been used to negotiate the exploration-exploitation tradeoff in multi-armed bandits [9, 32], learn algorithms which dynamically adjust their own learning rates [32, 33], and perform one-shot learning using external memory [27]. In the present work we explore the possibility of obtaining a causally-aware inner-loop learning algorithm.

5 Formalism for Causal Inference

5.1 Causal Bayesian Networks

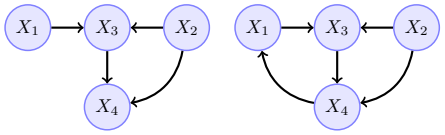

Figure 1: (a): Directed acyclic graph. The node X_3 is a collider on the path $X_1 \to X_3 \leftarrow X_2$ and a non-collider on the path $X_2 \to X_3 \to X_4$. (b): Cyclic graph obtained from (a) by adding a link from X_4 to X_1.

By combining graph theory and probability theory, the causal Bayesian network framework provides us with a graphical tool to formalize and test different levels of causal reasoning. This section introduces the main definitions underlying this framework and explains how to visually test for statistical independence [2, 4, 6, 7, 18, 21, 22, 23, 30].

A **graph** is a collection of nodes and links connecting pairs of nodes. The links may be directed or undirected, giving rise to **directed** or **undirected graphs** respectively.

A **path** from node X_i to node X_j is a sequence of linked nodes starting at X_i and ending at X_j. A **directed path** is a path whose links are directed and pointing from preceding towards following nodes in the sequence.

A **directed acyclic graph** is a directed graph with no directed paths starting and ending at the same node. For example, the directed graph in Fig. 1(a) is acyclic. The addition of a link from X_4 to X_1 gives rise to a cyclic graph (Fig. 1(b)).

A node X_i with a directed link to X_j is called **parent** of X_j. In this case, X_j is called **child** of X_i.

A node is a **collider** on a specified path if it has (at least) two parents on that path. Notice that a node can be a collider on a path and a non-collider on another path. For example, in Fig. 1(a) X_3 is a collider on the path $X_1 \to X_3 \leftarrow X_2$ and a non-collider on the path $X_2 \to X_3 \to X_4$.

A node X_i is an **ancestor** of a node X_j if there exists a directed path from X_i to X_j. In this case, X_j is a **descendant** of X_i.

A **graphical model** is a graph in which nodes represent random variables and links express statistical relationships between the variables.

A **Bayesian network** is a directed acyclic graphical model in which each node X_i is associated with the conditional distribution $p(X_i|\text{pa}(X_i))$, where $\text{pa}(X_i)$ indicates the parents of X_i. The joint distribution of all nodes in the graph, $p(X_{1:N})$, is given by the product of all conditional distributions, i.e. $p(X_{1:N}) = \prod_{i=1}^{N} p(X_i|\text{pa}(X_i))$.

When equipped with causal semantic, namely when describing the process underlying the data generation, a Bayesian network expresses both causal and statistical relationships among random variables—in such a case the network is called **causal**.

Assessing statistical independence in Bayesian networks. Given the sets of random variables $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, $\mathcal{X}$ and $\mathcal{Y}$ are statistically independent given $\mathcal{Z}$ if all paths from any element of $\mathcal{X}$ to any element of $\mathcal{Y}$ are **closed** (or **blocked**). A path is closed if at least one of the following conditions is satisfied:

(i) There is a non-collider on the path which belongs to the conditioning set $\mathcal{Z}$.
(ii) There is a collider on the path such that neither the collider nor any of its descendants belong to $\mathcal{Z}$.

5.2 An Intuitive Example of Cause-effect Reasoning in a CBN

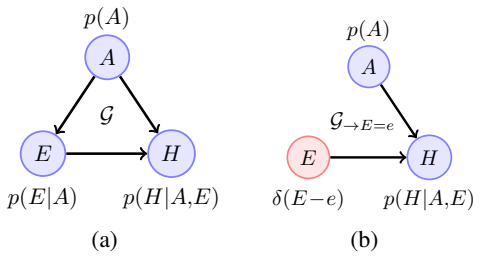

Figure 2: (a): A CBN $\mathcal{G}$ with a confounder for the effect of exercise (E) on heath (H) given by age (A). (b): Intervened CBN $\mathcal{G}_{\to E=e}$.

An example of CBN $\mathcal{G}$ is given in Fig. 2a, where E represents hours of exercise in a week, H cardiac health, and A age. Random variables are denoted by capital letters (e.g., E) and their values by small letters (e.g., e). The causal effect of E on H is the conditional distribution restricted to the path $E \to H$, i.e. excluding the path $E \leftarrow A \to H$. The variable A is called a *confounder*, as it confounds the causal effect with non-causal statistical influence.

Simply observing cardiac health conditioning on exercise level from $p(H|E)$ (associative reasoning) cannot answer if change in exercise levels cause changes in cardiac health (cause-effect reasoning), since there is always the possibility that correlation between the two is because of the common confounder of age.

The causal effect of $E = e$ can be seen as the conditional distribution $p_{\to E=e}(H|E=e)$[3] on the *intervened* CBN $\mathcal{G}_{\to E=e}$ resulting from replacing $p(E|A)$ with a delta distribution $\delta(E-e)$ (thereby removing the link from A to E) and leaving the remaining conditional distributions $p(H|E,A)$ and $p(A)$ unaltered (Fig. 2b). The rules of do-calculus [23, 24] tell us how to compute $p_{\to E=e}(H|E=e)$ using observations from $\mathcal{G}$. In this case $p_{\to E=e}(H|E=e) = \sum_A p(H|E=e,A)p(A)$[4]. Therefore, do-calculus enables us to reason in the intervened graph $\mathcal{G}_{\to E=e}$ even if our observations are from $\mathcal{G}$. This is the kind of causal reasoning possible in our observational data setting.

Such inferences are always possible if the confounders are observed, but in the presence of unobserved confounders, for many CBN structures the only way to compute causal effects is by collecting observations directly from the intervened graph, e.g. from $\mathcal{G}_{\to E=e}$ by fixing the value of the variable $E=e$ and observing the remaining variables—we call this process performing an actual intervention in the environment. In our interventional data setting the agent has access to such interventions.

5.3 Counterfactual Reasoning

Cause-effect reasoning can be used to correctly answer predictive questions of the type "Does exercising improve cardiac health?" by accounting for causal structure and confounding. However, it cannot answer retrospective questions about what *would have* happened. For example, given an individual i who has died of a heart attack, this method would not be able to answer questions of the type "What would the cardiac health of this individual have been had she done more exercise?". This type of question requires reasoning about a counterfactual world (that did not happen). To do this, we can first use the observations from the factual world and knowledge about the CBN to get an estimate of the specific latent randomness in the makeup of individual i (for example information about this specific patient's blood pressure and other variables as inferred by her having had a heart attack). Then, we can use this estimate to compute cardiac health under intervention on exercise. This procedure is called the *Abduction-Action-Prediction Method* [24] and is described below.

Assume, for example, the following model for $\mathcal{G}$ in Figure 2: $E = w_{AE}A + \eta$, $H = w_{AH}A + w_{EH}E + \epsilon$, where the weights w_{ij} represent the known causal effects in $\mathcal{G}$ and ϵ and η are terms of (e.g.) Gaussian noise that represent the latent randomness in the makeup of each individual. These noise variables are

[3] In the causality literature, this distribution would most often be indicated with $p(H|\text{do}(E=e))$. We prefer to use $p_{\to E=e}(H|E=e)$ to highlight that intervening on E results in changing the original distribution p, by structurally altering the CBN.

[4] Notice that conditioning on $E=e$ would instead give $p(H|E=e) = \sum_A p(H|E=e,A)p(A|E=e)$.

zero in expectation, so without access to their value for an individual we simply use $\mathcal{G}$: $E = w_{AE}A$, $H = w_{AH}A + w_{EH}E$ to make causal predictions. Suppose that for individual i we observe: $A = a^i$, $E = e^i$, $H = h^i$. We can answer the counterfactual question of "What if individual i had done more exercise, i.e. $E = e'$, instead?" by: a) *Abduction:* estimate the individual's specific makeup with $\epsilon^i = h^i - w_{AH}a^i - w_{EH}e^i$, b) *Action:* set E to more exercise e', c) *Prediction:* predict a new value for cardiac health as $h' = w_{AH}a^i + w_{EH}e' + \epsilon^i$.

6 RL Baselines

Figure 3: Reward distribution

We can also compare the performance of these agents to two standard model-free RL baselines. The Q-total Agent learns a Q-value for each action across all steps for all the episodes. The Q-episode Agent learns a Q-value for each action conditioned on the input at each time step $[o_t, a_{t-1}, r_{t-1}]$, but with no LSTM memory to store previous actions and observations. Since the relationship between action and reward is random between episodes, Q-total was equivalent to selecting actions randomly, resulting in a considerably negative reward (-1.247 ± 2.940). The Q-episode agent essentially makes sure to not choose the arm that is indicated by m_t to be the external intervention (which is assured to be equal to -5), and essentially chooses randomly otherwise, giving a reward close to 0 (0.080 ± 2.077).

7 Additional Experiments

The purview of the previous experiments was to show a proof of concept on a simple tractable system, demonstrating that causal induction and inference can be learned and implemented via a meta-learned agent. In the following, we additionally demonstrate counterfactual reasoning, and scale up our results to more complex systems in two new experiments.

7.1 Experiment 3: Counterfactual Setting

In Experiment 3, the agent was again allowed to make interventions as in Experiment 2, but in this case the quiz phase task entailed answering a counterfactual question. We explain here what a counterfactual question in our experimental domain looks like. Assume $X_i = \sum_j w_{ji}X_j + \epsilon_i$ where ϵ_i is distributed as $\mathcal{N}(0.0, 0.1)$ (giving the conditional distribution $p(X_i|\text{pa}(X_i)) = \mathcal{N}(\sum_j w_{ji}X_j, 0.1)$ as described in Section 3). After observing the nodes $X_{2:N}$ (X_1 is hidden) in the CBN in one sample, we can infer this latent randomness ϵ_i for each observable node X_i (i.e. *abduction*) and answer counterfactual questions like "What would the values of the nodes be, had X_i instead taken on a different value than what we observed?", for any of the observable nodes X_i. We test three new agents, two of which are learned: "Active Counterfactual", "Random Counterfactual", and "Optimal Counterfactual Baseline" (not learned).

Counterfactual Agents: This agent is the same as the Interventional agent, but trained on tasks in which the latent randomness in the last information phase step $t = T - 1$ (where some $X_p = +5$) is stored and the same randomness is used in the quiz phase step $t = T$ (where some $X_f = -5$). While the question our agents have had to answer correctly so far in order to maximize their reward in the quiz phase was "Which of the nodes $X_{2:N}$ will have the highest value when X_f is set to -5?", in this setting, we ask "Which of the nodes $X_{2:N}$ would have had the highest value in the last step of the information phase, if instead of having the intervention $X_p = +5$, we had the intervention $X_f = -5$?". We run active and random versions of this agent as described in the main text.

Optimal Counterfactual Baseline: This baseline receives the true CBN and does exact abduction of the latent randomness based on observations from the penultimate step of the information phase, and combines this correctly with the appropriate interventional inference on the true CBN in the quiz phase.

Results

We focus on two key questions in this experiment.

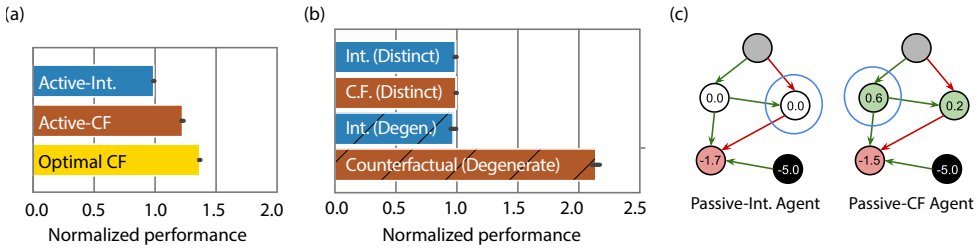

Figure 4: Experiment 3. Agents do counterfactual reasoning. a) Performance of the agents tested in this experiment. Note that performance can be above 1.0 since the counterfactual agent can theoretically perform better than the optimal interventional baseline, which doesn't have access to noise information. See main text for details. b) Performance split by if the maximum node value in the quiz phase is degenerate (Deg.) or distinct (Dist.). c) Quiz phase for an example test-CBN. See Figures in Main text for a legend. Here, the left panel shows $\mathcal{G}_{\rightarrow X_j=-5}$ and the nodes taking the mean values prescribed by $p_{\rightarrow X_j=-5}(X_{1:N\setminus j}|X_j=-5)$. We see that the Active-Int. Agent's choice is consistent with maximizing on these node values, where it makes a random choice between two nodes with the same value. The right panel panel shows $\mathcal{G}_{\rightarrow X_j=-5}$ and the nodes taking the exact values prescribed by the means of $p_{\rightarrow X_j=-5}(X_{1:N\setminus j}|X_j=-5)$, combined with the specific randomness inferred from the previous time step. As a result of accounting for the randomness, the two previously degenerate maximum values are now distinct. We see that the Active-CF. agent's choice is consistent with maximizing on these node values.

(i) Do our agents learn to do counterfactual inference? The Active-Counterfactual Agent achieves higher performance than the maximum possible performance using only causal reasoning (Figure 4a). This indicates that the agent learns to infer and apply noise information from the last step of the information phase. To evaluate whether this difference is driven by the agent's use of abduction, we split the test set into two groups, depending on whether or not the decision for which node will have the highest value in the quiz phase is affected by the latent randomness, i.e. whether or not the node with the maximum value in the quiz phase changes if the noise is resampled. This is most prevalent in cases where the maximum expected reward is degenerate, i.e. where several nodes give the same maximum reward (denoted by hatched bars in Figure 4b). Here, agents with no access to the randomness have no basis for choosing one over the other, but different noise samples can give rise to significant differences in the actual values that these degenerate nodes have. We see indeed that there is no difference in the rewards received by the Active-Counterfactual and Active-Interventional Agents in the cases where the maximum values are distinct, however the Active-Counterfactual Agent significantly outperforms the Active-Interventional Agent in cases where there are degenerate maximum values. This performance increase is very high since in most cases where the maximum values are degenerate, this maximum value is close to 0.0. Thus, taking the noise into account gives the Counterfactual agent a huge relative advantage in these cases.

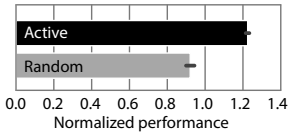

Figure 5: Active and Random Counterfactual Agents

(ii) Do our agents learn to make useful interventions in the service of a counterfactual task? The Active-Counterfactual Agent's performance is significantly greater than the Random-Counterfactual Agent's (Fig. 5). This indicates that when the agent is allowed to choose its actions, it makes tailored, non-random choices about the interventions it makes and the data it wants to observe – even in the service of a counterfactual objective.

7.2 Experiment 4: Non-linear Causal Graphs

In this experiment, we generalize some of our results to nonlinear, non-Gaussian causal graphs which are more typical of real-world causal graphs and to demonstrate that our results hold without loss of generality on such systems.

Here we investigate causal Bayesian networks (CBNs) with a quadratic dependence on the parents by changing the conditional distribution to $p(X_i|\text{pa}(X_i))=\mathcal{N}(\frac{1}{N_i}\sum_j w_{ji}(X_j+X_j^2),\sigma)$. Here, although

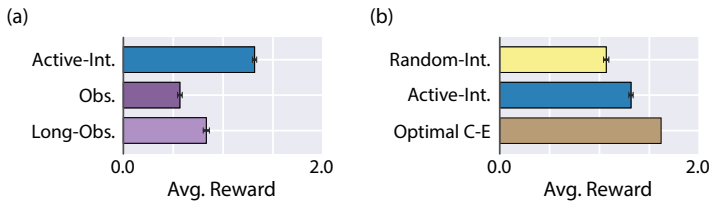

Figure 6: Results for non-linear graphs. (a) Comparing average episode reward for agents trained with different data. (b) Comparing information phase intervention policies.

each node is normally distributed given its parents, the joint distribution is not multivariate Gaussian due to the non-linearity in how the means are determined. We find that the Long-Observational Agent achieves more reward than the Observational Agent indicating that the agent is in fact learning the statistical dependencies between the nodes, within an episode. [5] We also find that the Active-Interventional Agent achieves reward well above the best agent with access to only observational data (Long-Observational in this case) indicating an ability to reason from interventions. We also see that the Active-Interventional Agent performs better than the Random-Interventional Agent, indicating an ability to choose informative interventions.

7.3 Experiment 5: Larger Causal Graphs

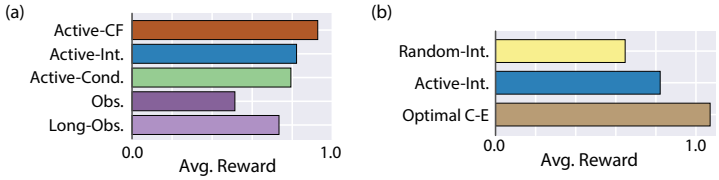

Figure 7: Results for $N = 6$ graphs. (a) Comparing average episode reward for agents trained with different data. (b) Comparing information phase intervention policies.

In this experiment we scaled up to larger graphs with $N = 6$ nodes, which afforded considerably more unique CBNs than with $N = 5$ (1.4×10^7 vs 5.9×10^4). As shown in Fig. 7a, we find the same pattern of behavior noted in the main text where the rewards earned are ordered such that Observational agent $<$ Active-Conditional agent $<$ Active-Interventional agent $<$ Active-Counterfactual agent. We see additionally in Fig. 7b that the Active-Interventional agent performs significantly better than the baseline Random-Interventional agent, indicating an ability to choose non-random, informative interventions.

References

[1] M. Andrychowicz, M. Denil, S. Gomez, M. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. D. Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pages 3981–3989, 2016.

[2] D. Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[3] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] The conditional distribution $p(X_{1:N \setminus j} | X_j = 5)$, and therefore Conditional Agents, were non-trivial to calculate for the quadratic case, and was thus omitted.

[5] N. Bramley, A. Rothe, J. Tenenbaum, F. Xu, and T. Gureckis. Grounding compositional hypothesis generation in specific instances. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 2018.

[6] S. Chiappa and W. S. Isaac. *A Causal Bayesian Networks Viewpoint on Fairness*, volume 547 of *IFIP AICT*, pages 3–20. Springer Nature Switzerland, 2019.

[7] P. Dawid. Fundamentals of statistical causality. Technical report, University Colledge London, 2007.

[8] M. Denil, P. Agrawal, T. D. Kulkarni, T. Erez, P. Battaglia, and N. de Freitas. Learning to perform physics experiments via deep reinforcement learning. *arXiv preprint arXiv:1611.01843*, 2016.

[9] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

[10] P. M. Fernbach, A. Darlow, and S. A. Sloman. Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, 21(3):329–336, 2010.

[11] P. M. Fernbach and B. Rehder. Cognitive shortcuts in causal inference. *Argument & Computation*, 4(1):64–88, 2013.

[12] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

[13] N. D. Goodman, T. D. Ullman, and J. B. Tenenbaum. Learning a theory of causality. *Psychological review*, 118(1):110, 2011.

[14] E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

[15] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[18] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[19] J. L. McClelland, M. M. Botvinick, D. C. Noelle, D. C. Plaut, T. T. Rogers, M. S. Seidenberg, and L. B. Smith. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, 14(8):348–356, 2010.

[20] V. Mnih, A. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783, 2016.

[21] K. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

[22] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[23] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[24] J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

[25] B. Rehder. Independence and dependence in human causal reasoning. *Cognitive psychology*, 72:54–107, 2014.

[26] B. Rehder and M. R. Waldmann. Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & cognition*, 45(2):245–260, 2017.

[27] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016.

[28] L. Schulz. Finding new facts; thinking new thoughts. In *Advances in child development and behavior*, volume 43, pages 269–294. Elsevier, 2012.

[29] H. T. Siegelmann and E. D. Sontag. On the computational power of neural nets. *Journal of computer and system sciences*, 50(1):132–150, 1995.

[30] P. Spirtes, C. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000.

[31] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[32] J. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *CoRR*, abs/1611.05763, 2016.

[33] J. X. Wang, Z. Kurth-Nelson, D. Kumaran, D. Tirumala, H. Soyer, J. Z. Leibo, D. Hassabis, and M. Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860, 2018.