Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning

Sumedh A Sontakke [1] **Arash Mehrjou** [2] **Laurent Itti** [1] **Bernhard Schölkopf** [2]

Abstract

Animals exhibit an innate ability to learn regularities of the world through interaction. By performing experiments in their environment, they are able to discern the causal factors of variation and infer how they affect the world's dynamics. Inspired by this, we attempt to equip reinforcement learning agents with the ability to perform experiments that facilitate a categorization of the rolled-out trajectories, and to subsequently infer the causal factors of the environment in a hierarchical manner. We introduce *causal curiosity*, a novel intrinsic reward, and show that it allows our agents to learn optimal sequences of actions and discover causal factors in the dynamics of the environment. The learned behavior allows the agents to infer a binary quantized representation for the ground-truth causal factors in every environment. Additionally, we find that these experimental behaviors are semantically meaningful (e.g., our agents learn to lift blocks to categorize them by weight), and are learnt in a self-supervised manner with approximately 2.5 times less data than conventional supervised planners. We show that these behaviors can be re-purposed and fine-tuned (e.g., from lifting to pushing or other downstream tasks). Finally, we show that the knowledge of causal factor representations aids zero-shot learning for more complex tasks.

1. Introduction

Discovering causation in environments an agent might encounter remains an open and challenging problem for reinforcement learning (Bengio et al., 2013; Schölkopf, 2015). In physical systems, causal factors such as gravity or friction affect the outcome of behaviors an agent might perform. Thus, there has been recent interest in attempting to train agents to be robust or invariant against varying values of such causal factors, allowing them to learn modular behaviors that are useful across tasks. Most model-based approaches take the form of Bayes Adaptive Markov Decision Processes (BAMDPs) (Zintgraf et al., 2019) or Hidden Parameter MDPs (Hi-Param MDPs) (Doshi-Velez & Konidaris, 2016; Yao et al., 2018; Killian et al., 2017; Perez et al., 2020) which condition the transition and/or reward function of each environment on hidden parameters.

Formally, let $\mathbf{s} \in \mathcal{S}$, $\mathbf{a} \in \mathcal{A}$, $\mathbf{r} \in \mathcal{R}$, $\mathbf{h} \in \mathcal{H}$ where $\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$, and $\mathcal{H}$ are the set of states, actions, rewards and admissible causal factors, respectively. In the physical world, examples of the parameter $\mathbf{h}_j \in \mathcal{H}$ might include gravity, coefficients of friction, masses and sizes of objects. Hi-Param MDP or BAMDP approaches treat each $\mathbf{h}_j \in \mathcal{H}$ as a latent variable for which an embedding is learnt during training (often using variational methods (Kingma et al., 2014; Ilse et al., 2019)). Let $\mathbf{a}_{0:T}$ be a sequence of actions taken by an agent to maximize an external reward resulting in a state trajectory $\mathbf{s}_{0:T}$. The above approaches define a probability distribution over the entire observable sequence (i.e., rewards, states, actions) as $p(\mathbf{r}_{0:T}, \mathbf{s}_{0:T}, \mathbf{a}_{0:T-1})$ which factorizes as

$$\prod_{t=1}^{T-1} p(\mathbf{r}_{t+1}|\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, \mathbf{z}) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t, \mathbf{z}) p(\mathbf{a}_t|\mathbf{s}_t, \mathbf{z})$$

conditioned on the latent variable $\mathbf{z}$, a representation for the unobserved causal factors. At test time, in a new environment, the agent infers $\mathbf{z}$ by observing the trajectories produced by its initial actions issued by the latent conditioned policy obtained during training.

In practice, discovering causal factors in a physical environment is prone to various challenges that are due to the disjointed nature of the influence of these factors on the produced trajectories. More specifically, at each time step, the transition function is affected by a subset of global causal factors. This subset is implicitly defined on the basis of the current state and the action taken. For example, if a body in an environment loses contact with the ground, the coefficient of friction between the body and the ground no longer affects the outcome of any action that is taken. Likewise, the outcome of an upward force applied by the agent to a body on the ground is unaffected by the friction coefficient.

[1]University of Southern California [2]Max Planck Institute for Intelligent Systems. Correspondence to: Sumedh A Sontakke <ssontakk@usc.edu>.

Pre-print. Do not distribute.

Without knowledge of how independent causal mechanisms affect the outcome of a particular action in a given state in an environment, it becomes impossible for the agent to conclude where an encountered variation came from. Unsurprisingly, Hi-Param and BAMDP approaches fail to learn a disentangled embedding of the causal factors, making their behaviors uninterpretable. For example, if, in an environment, a body remains stationary under a particular force, the Hi-Param or BAMDP agent may apply a higher force to achieve its goal of perhaps moving the body, but will be unable to conclude whether the "un-movability" was caused by a high friction coefficient, or high mass. Additionally, these approaches require substantial reward engineering, making it difficult to apply them outside the simulated environments they are tested in.

Our goal is, instead of focusing on maximizing reward for a particular task, to allow agents to discover causal processes through exploratory interaction. During training, our agents discover self-supervised experimental behaviors which they apply to a set of training environments. These behaviors allow them to learn about the various causal mechanisms that govern the transitions in each environment. During inference in a novel environment, they perform these discovered behaviors sequentially and use the outcome of each behavior to infer the embedding for a single causal factor (Figure 1), allowing us to recover a disentangled embedding describing the causal factors of an environment.

The main challenge while learning a disentangled representation for the causal factors of the world is that several causal factors may affect the outcome of behaviors in each environment. For example, when pushing a body on the ground, the outcome, i.e., whether the body moves, or how far the body is pushed, depends on several factors, e.g., mass, shape and size, frictional coefficients, etc. However, if, instead of pushing on the ground, the agent executes a perfect grasp-and-lift behavior, only mass will affect whether the body is lifted off the ground or not.

Thus, it is clear that not all experimental behaviors are created equal and that the outcomes of some behaviors are caused by fewer causal factors than others. Our agents learn these behaviors without supervision using *causal curiosity*, an intrinsic reward. The outcome of a single such experimental behavior is then used to infer a binary quantized embedding describing the single isolated causal factor. While causal factors of variation in a physical world are easily identifiable to humans, a concrete definition is required to formalize our proposed method.

Definition 1 (Causal factors)**.** *Consider the POMDP ($\mathcal{O}$, $\mathcal{S}$, $\mathcal{A}$, ϕ, θ, r) with observation space $\mathcal{O}$, state space $\mathcal{S}$, action space $\mathcal{A}$, the transition function ϕ, emission function θ, and the reward function r. Let $\mathbf{o}_{0:T} \in \mathcal{O}^T$ denote a trajectory of observations of length T. Let $d(\cdot,\cdot) : \mathcal{O}^T \times \mathcal{O}^T \to \mathbb{R}_+$ be a distance function defined on the space of trajectories of length T. The set $H = \{\mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_{K-1}\}$ is called a set of $\epsilon-$causal factors if for every $\mathbf{h}_j \in H$, there exists a unique sequence of actions $\mathbf{a}_{0:T}$ that clusters the observation trajectories into m disjoint sets $C_{1:m}$ such that $\forall C_a, C_b$, a minimum separation distance of ϵ is ensured:*

$$\min\{d(\mathbf{o}_{0:T}, \mathbf{o}'_{0:T}) : \mathbf{o}_{0:T} \in C_a, \mathbf{o}'_{0:T} \in C_b\} > \epsilon \quad (1)$$

and that $\mathbf{h}_j$ is the cause of the obtained trajectory of states i.e. $\forall v \neq v'$,

$$p(\mathbf{o}_{0:T}|do(\mathbf{h}_j = v), \mathbf{a}_{0:T}) \neq p(\mathbf{o}_{0:T}|do(\mathbf{h}_j = v'), \mathbf{a}_{0:T}) \quad (2)$$

where $do(\mathbf{h}_j)$ corresponds to an intervention on the value of the causal factor $\mathbf{h}_j$.

According to Def. 1, a causal factor $\mathbf{h}_j$ is a variable in the environment the value of which, when intervened on (i.e., varied) using $do(\mathbf{h}_j)$ over a set of values, results in trajectories of observations that are divisible into disjoint clusters $C_{1:m}$ under a particular sequence of actions $\mathbf{a}_{0:T}$. These clusters represent the quantized values of the causal factor. For example, mass, which is a causal factor of a body, under an action sequence of a grasping and lifting motion, may result in 2 clusters, liftable (low mass) and not-liftable (high mass).

However, such an action sequence is not known in advance. Therefore, discovering a causal factor in the environment boils down to finding a sequence of actions that makes the effect of that factor prominent by producing clustered trajectories for different values of that environmental factor. For simplicity, here we assume binary clusters. For a gentle introduction to the intuition about this definition, we refer the reader to Appendix E. For an introduction to causality and $do(\cdot)$ notation, see (Pearl, 2009; Spirtes, 2010; Schölkopf, 2019; Elwert, 2013).

Our contributions of our work are as follows:

- **Causal POMDPs:** We extend Partially Observable Markov Decision Processes (POMDPs) by explicitly modelling the effect of causal factors on observations.
- **Unsupervised Behavior:** We equip agents with the ability to perform experiments and behave in a semantically meaningful manner in a set of environments in an unsupervised manner. These behaviors can expose or obfuscate specific independent causal mechanisms that occur in the world surrounding the agent, allowing the agent to "experiment" and learn.
- **Disentangled Representation Learning:** We introduce an minimalistic intrinsic reward, *causal curiosity*, which allows our agents to discover these behaviors

without human-engineered complex rewards. The outcomes of the experiments are used to learn a disentangled quantized binary representation for the causal factors of the environment, analogous to the human ability to conclude whether objects are light/heavy, big/small etc.

- **Sample Efficiency:** Through extensive experiments, we conclude that knowledge of the causal factors aids sample efficiency in two ways - first, that the knowledge of the causal factors aids transfer learning across multiple environments; second, the learned experimental behaviors can be re-purposed for downstream tasks.

2. Method

Consider a set of N environments $\mathcal{E}$ with $\mathbf{e}^i \in \mathcal{E}$ where $\mathbf{e}^i$ denotes the i^{th} environment. Each causal factor $\mathbf{h}_j \in H$ is itself a random variable which assumes a particular value for every instantiation of an environment. Thus, every environment $\mathbf{e}^i$ is represented with the values assumed by its causal factors $\{\mathbf{h}^i_j, j = 0, 1, \ldots, K-1\}$. For each environment $\mathbf{e}^i$, $(\mathbf{z}^i_0, \mathbf{z}^i_1 ... \mathbf{z}^i_{K-1})$ represents the disentangled embedding vector corresponding to the physical causal factors where $\mathbf{z}^i_j$ encodes $\mathbf{h}^i_j$.

2.1. POMDP Setup

2.1.1. Classical POMDPs

Classical POMDPs ($\mathcal{O}$, $\mathcal{S}$, $\mathcal{A}$, ϕ, θ, r) consist of an observation space $\mathcal{O}$, state space $\mathcal{S}$, action space $\mathcal{A}$, the transition function ϕ, emission function θ, and the reward function r. An agent in an unobserved state $\mathbf{s}_t$ takes an action $\mathbf{a}_t$ and consequently causes a transition in the environment through $\phi(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$. The agent receives an observation $\mathbf{o}_{t+1} = \theta(\mathbf{s}_{t+1})$ and a reward $\mathbf{r}_{t+1} = r(\mathbf{s}_t, \mathbf{a}_t)$.

2.1.2. Causal POMDPs

Our work divides the unobserved state $\mathbf{s} \in \mathcal{S}$ at each timestep into two portions - the *controllable state* $\mathbf{s}^c$ and the *uncontrollable state* $\mathbf{s}^u$. The uncontrollable portion of the state $\mathbf{s}^u$ consists of the causal factors of the environment. We assume that these remain constant during the interaction of the agent with a single instance of the environment. For example, the value of the gravitational acceleration does not change for a single environment. For the following discussion, we refer to the uncontrollable state as causal factors as in Def 1 i.e., $\mathbf{s}^u = \mathcal{H}$.

The controllable state $\mathbf{s}^c$ consists of state variables such as positions and orientations of objects, location of end-effectors of the agent etc. Thus, by executing particular action sequences the agent can manipulate this portion of the state, which is hence controllable by the agent.

2.1.3. Transition Probability

A trajectory of the controllable state is dependent on both the action sequence that the agent executes and a subset of the causal factors. At each time step, only a subset of the causal factors of an environment affect the transition in the environment. This subset is implicitly selected by the employed policy for every state of the trajectory (depicted as a Gated Causal Graph (Figure 2)). For example, the outcome of an upward force applied by the agent to a body on the ground is unaffected by the friction coefficient between the body and the ground.

Thus the transition function of the controllable state is:

$$\phi(\mathbf{s}^c_{t+1}|\mathbf{s}^c_t, f_{sel}(\mathcal{H}, \mathbf{s}^c_t, \mathbf{a}_t), \mathbf{a}_t) \tag{3}$$

where f_{sel} is the implicit Causal Selector Function which selects the subset of causal factors affecting the transition defined as:

$$f_{sel} : \mathcal{H} \times \mathcal{S} \times \mathcal{A} \to \wp(\mathcal{H}) \tag{4}$$

where $\wp(\mathcal{H})$ is power-set of $\mathcal{H}$ and $f_{sel}(\mathcal{H}, \mathbf{s}^c_t, \mathbf{a}_t) \subset \mathcal{H}$ is the set of effective causal factors for the transition $\mathbf{s}_t \to \mathbf{s}_{t+1}$ i.e., $\forall v \neq v'$ and $\forall \mathbf{h}_j \in f_{sel}(\mathcal{H}, \mathbf{s}^c_t, \mathbf{a}_t)$:

$$\phi(\mathbf{s}^c_{t+1}|do(\mathbf{h}_j = v), \mathbf{s}^c_t, \mathbf{a}_t) \neq \phi(\mathbf{s}^c_{t+1}|do(\mathbf{h}_j = v'), \mathbf{s}^c_t, \mathbf{a}_t) \tag{5}$$

where $do(\mathbf{h}_j)$ corresponds to an external intervention on the factor $\mathbf{h}_j$ in an environment.

Intuitively, this means that if an agent takes an action $\mathbf{a}_t$ in the controllable state $\mathbf{s}^c_t$, the transition to $\mathbf{s}^c_{t+1}$ is caused by a subset of the causal factors $f_{sel}(\mathcal{H}, \mathbf{s}^c_t, \mathbf{a}_t)$. For example, if a body on the ground (i.e., state $\mathbf{s}^c_t$) is thrown upwards (i.e., action $\mathbf{a}_t$), the outcome $\mathbf{s}_{t+1}$ is caused by the causal factor gravity (i.e., $f_{sel}(\mathcal{H}, \mathbf{s}^c_t, \mathbf{a}_t) = \{\text{gravity}\}$), a singleton subset of the global set of causal factors. The $do()$ notation expresses this causation. If an external intervention on a causal factor is performed, e.g., if somehow the value of gravity was changed from v to v', the outcome of throwing the body up from the ground, $\mathbf{s}_{t+1}$, would be different.

2.1.4. Emission Probability

The agent neither has access to the controllable state, nor to the causal factors of each environment. It receives an observation described by the function:

$$\mathbf{o}_{t+1} = \theta(\mathbf{s}^c_t, f_{sel}(\mathcal{H}, \mathbf{s}^c_t, \mathbf{a}_t), \mathbf{a}_t) \tag{6}$$

where f_{sel} is the implicit Causal Selector Function.

2.2. Training the Experiment Planner

The agent has access to a set of training environments with multiple causal factors varying simultaneously. Our goal

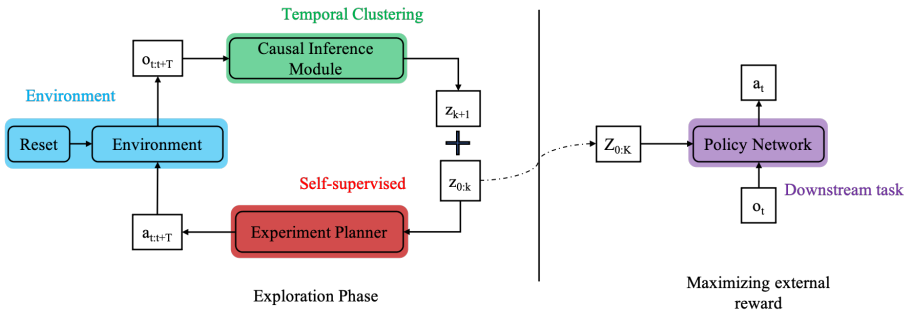


Figure 1. Overview of Inference. The exploration loop produces a series of K experiments allowing the agent to infer the representations for K causal factors. After exploration, the agent utilizes the acquired knowledge for downstream tasks. The details for the inference procedure are provided in Supplementary Material Algorithm 2.

is to allow the agent to discover action sequences $\mathbf{a}_{0:T}$ such that the resultant observation trajectory $\mathbf{o}^i_{0:T}$ is caused by a single causal factor i.e., $\forall t < T, f_{sel}(\mathcal{H}, \mathbf{s}^c_t, \mathbf{a}_t) =$ constant and $|f_{sel}(\mathcal{H}, \mathbf{s}^c_t, \mathbf{a}_t)| = 1$. Consequently, $\mathbf{o}^i_{0:T}$ can be used to learn a representation $\mathbf{z}^i_j$ for the causal factor $f_{sel}(\mathcal{H}, \mathbf{s}^c_t, \mathbf{a}_t)$ for each environment $\mathbf{e}^i$.

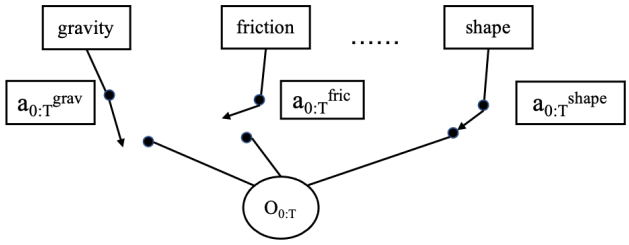


Figure 2. Gated Causal Graph. A subset of the unobserved parent causal variables influence the observed variable $\mathbf{O}$. The action sequence $\mathbf{a}_{0:T}$ serves a gating mechanism, allowing or blocking particular edges of the causal graph using the implicit Causal Selector Function (Equation 4).

We motivate this from the perspective of algorithmic information theory (Janzing & Schölkopf, 2010). Consider the Gated Directed Acyclic Graph of the observed variable $\mathbf{O}$ and its causal parents (Figure 2). Each causal factor has its own causal mechanism, jointly bringing about $\mathbf{O}$. A central assumption of our approach is that causal factors are independent, i.e., the Independent Mechanisms Assumption (Schölkopf et al., 2012; Parascandolo et al., 2018; Schölkopf, 2019). The information in $\mathbf{O}$ is then the sum of information "injected" into it from the multiple causes, since, loosely speaking, for information to cancel, the mechanisms would need to be algorithmically dependent (Janzing & Schölkopf, 2010). Intuitively, the information content in $\mathbf{O}$ will be greater for a larger number of causal parents in the graph. Interestingly, a similar argument has been made to justify the thermodynamic arrow of time (Chaves et al., 2014; Janzing et al., 2016): while a microscopic time evolution is invertible, the assumption of algorithmic independence for initial conditions and forward mechanisms generates an asymmetry. To invert time, the backward mechanism would need to depend on the initial state.

Thus we attempt to find an action sequence $\mathbf{a}_{0:T}$ for which the number of causal parents of the resultant observation $\mathbf{O}$ is low, i.e., the complexity of $\mathbf{O}$ is low. One could conceive of this by assuming that the generative model for $\mathbf{O}$, $\mathbf{M}$ has low Kolmogorov Complexity. Here, a low capacity bi-modal model is assumed. We utilize Minimum Description Length $L(\cdot)$ (MDL) as a tractable substitute of the Kolmogorov Complexity (Rissanen, 1978; Grunwald, 2004)).

Causal curiosity solves the following optimization problem.

$$\mathbf{a}^*_{0:T} = \arg\min_{\mathbf{a}_{0:T}} (L(\mathbf{M}) + L(\mathbf{O}|\mathbf{M})) \tag{7}$$

where each observed trajectory $\mathbf{O} = \mathbf{O}(\mathbf{a}_{0:T})$ is a function of the action sequence. As mentioned earlier, the model is fixed in this formulation; hence, the first term $L(\mathbf{M})$ is constant and not a function of the actions. The MDL of the trajectories given binary categorization model, $-L(\mathbf{O}|\mathbf{M})$, is the inherent reward function that is fed back to the RL agent. We regard this reward function as *causal curiosity*. See Supplementary Material B for implementation details.

2.3. Causal Inference Module

By maximizing the causal curiosity reward it is possible to achieve behaviors which result in trajectories of states only caused by a single hidden parameter. Subsequently, we utilize the outcome of performing these experimental behaviors in each environment to infer a representation for the causal factor isolated by the experiment in question.

We achieve this through clustering. An action sequence $\mathbf{a}_{0:T} \sim \text{CEM}(\cdot|\mathbf{z}^i_{0:j-1})$ is sampled from the Model Predictive Control Planner (Camacho & Alba, 2013) and applied to each of the training environments. The learnt clustering

Algorithm 1 Recursive Training Scheme

Initialize $j = 0$
Initialize training environment set $Envs$
function Train(j, $\mathbf{z}^i_{0:j}$, $Envs$)
  if $j == K$ **then**
    Return
  end if
  for iteration m to M **do**
    Sample experimental behavior $\mathbf{a}_{0:T} \sim$ CEM($\cdot|\mathbf{z}^i_{0:j}$)
    for i^{th} env in $Envs$ **do**
      Apply $\mathbf{a}_{0:T}$ to env
      Collect $\mathbf{O}^i = \mathbf{o}^i_{0:T}$
      Reset env
    end for
    Calculate $-L(\mathbf{O}|\mathbf{M})$ given that $\mathbf{M}$ is bimodal clustering model
    Update CEM($\cdot$) distribution with highest reward trajectories
  end for
  Use learnt $q_M(\mathbf{z}^i_j|\mathbf{O}, \mathbf{z}^i_{0:j})$ for cluster assignment of each env in $Envs$ i.e. $\mathbf{z}^i_j = q_M(\mathbf{z}|\mathbf{O}^i, \mathbf{z}^i_{0:j})$
  Update $j = j + 1$
  Train(j, $\mathbf{z}^i_{0:j}$, $Envs = \{\mathbf{e}^i : \mathbf{z}^i_{j-1} = 0\}$)
  Train(j, $\mathbf{z}^i_{0:j}$, $Envs = \{\mathbf{e}^i : \mathbf{z}^i_{j-1} = 1\}$)
end function

model $\mathbf{M}$ is then used to infer a representation for each environment using the collected outcome $\mathbf{O}^i$ obtained by applying $\mathbf{a}_{0:T}$ to each environment.

$$\mathbf{z}^i_j = q_M(\mathbf{z}|\mathbf{O}^i, \mathbf{z}^i_{0:j-1}) \tag{8}$$

The learnt representation $\mathbf{z}$ is the cluster membership obtained from the learnt clustering model $\mathbf{M}$. It is binary in nature. This corresponds to the quantization of the continuous spectrum of values a causal factor takes in the training set into high and low values.

2.4. Interventions on beliefs

Having learnt about the effects of a single causal factor of the environment we wish to learn such experimental behaviors for each of the remaining hidden parameters that may vary in the training environments. To achieve this, in an ideal setting, the agent would require access to the generative mechanism of the environments it encounters. Ideally, it would hold the values of the causal factor already learnt about a constant i.e. $do(\mathbf{h}_j = constant)$, and intervene over (vary the value of) another causal factor over a set of values V i.e. $do(\mathbf{h}_{j'} = v)$ such that $v \in V$. For example, if a human scientist were to study the effects of a causal factor, say mass of a body, they would hold the values of all other causal factors constant (e.g., interact with cubes of the same size and external texture), and vary only mass to see how it affects the outcome of specific behaviors they apply to each body.

However, in the real world the agent does not have access to the generative mechanism of the environments it encounters, but merely has the ability to act in them. Thus, it can intervene on the representations of a causal factor of the environment i.e. $do(\mathbf{z}_i = constant)$. For, example having learnt about gravity, the agent picks all environments it believes have the same gravity, and uses them to learn about a separate causal factor say, friction.

Thus, to learn about the j^{th} causal factor, the agent proceeds in a tree-like manner, dividing each of the 2^{j-1} clusters of training environments into two sub-clusters corresponding to the binary quantized values of the j^{th} causal factor. Each level of this tree corresponds to a single causal factor.

$$Envs = \{\mathbf{e}^i : \mathbf{z}^i_{j-1} = k\}, k \in \{0, 1\} \tag{9}$$

This process continues iteratively (Algorithm 1 and Figure 3), where for each cluster of environments, a new experiment learns to split the cluster into 2 sub-clusters depending on the value of a hidden parameter. At level n, the agent produces 2^n experiments, having already intervened on the binary quantized representations of n causal factors.

3. Related Work

Curiosity for robotics is not a new area of research. Pioneered by Schmidhuber in the 1990s, (Schmidhuber, 1991b;a; 2006; 2010), (Ngo et al., 2012), (Pathak et al., 2017) curiosity is described as the motivation behind the behavior of an agent in an environment for which the outcome is unpredictable, i.e., an intrinsic reward that motivates the agent to explore the unseen portions of the state space (and subsequent transitions). We refer the reader to Supplementary Material F for a more exhaustive review of related literature.

4. Experiments

Our work has 2 main thrusts - the discovered *experimental behaviors* and the *representations* obtained from the outcome of the behaviors in environments. We visualize these learnt behaviors and verify that they are indeed semantically meaningful and interpretable. We quantify the utility of the learned behaviors by using the behaviors as pre-training for a downstream task. In our experimental setup, we verify that these behaviors are indeed invariant to all other causal factors except one.
We visualize the representations obtained using these behaviors and verify that they are indeed the binary quantized representations for each of the ground truth causal factors that we manipulated in our experiments. Finally, we ver-

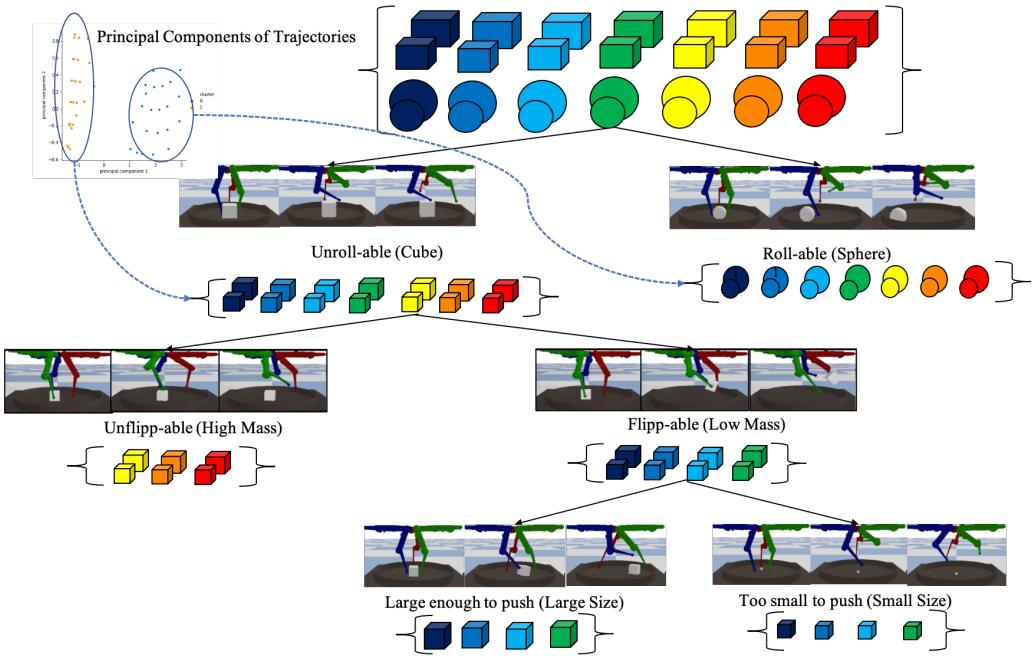


Figure 3. Discovered hierarchical latent space. The agent learns experiments that differentiate the full set of blocks in `ShapeSizeMass` into hierarchical binary clusters. At each level, the environments are divided into 2 clusters on the basis of the value of a single causal factor. We also show the principal components of the trajectories in the top left. For brevity, the full of extent of the tree is not depicted here. For each level of hierarchy k, there are 2^k number of clusters.

ify that the knowledge of the representation does indeed aid transfer learning and zero-shot generalizability in downstream tasks.

Causal World. We use the Causal World Simulation (Ahmed et al., 2020) based on the Pybullet Physics engine to test our approach. The simulator consists of a 3-fingered robot, with 3 joints on each finger. We constrain each environment to consist of a single object that the agent can interact with. The causal factors that we manipulate for each of the objects are size, shape and mass of the blocks. The simulator allows us to capture and track the positions and velocities of each of the movable objects in an environment.

4.1. Visualizing Discovered Behaviors

We would like to analyze whether the discovered experimental behaviors are human interpretable, i.e., *are the experimental behaviors discovered in each of the setups semantically meaningful?* We find that our agents learn to perform several useful behaviors without any supervision. For instance, to differentiate between objects with varying mass, we find that they acquire a perfect grasp-and-lift behavior with an upward force. In other random seed experiments, the agents learn to lift the blocks by using the wall of the environment for support. To differentiate between cubes and spheres, the agent discovers a pushing behavior which gently rolls the spheres along a horizontal direction. Qualitatively, we find that these behaviors are stable and predictable. See videos of discovered behaviors `here` (website under construction).

Concurrent with the objective they are trained on, we find that the acquired behaviors impose structure on the outcome when applied to each of the training environments. The outcome of each experimental behavior on the set of training environments results in dividing it into 2 subsets corresponding to the binary quantized values of a single factor, e.g., large or small, while being invariant to the values of other causal factors of the environments. We also perform ablation studies where instead of providing the full state vector, we provide only one coordinate (e.g., only x, y or z coordinate of the block). We find that causal curiosity results in behaviors that differentiate the environments based on outcomes along the direction provided. For example, when only the x coordinate was provided, the agent learned to evaluate mass by applying a pushing behavior along the x direction. Similarly, a lifting behavior was obtained when only the z coordinate was supplied to the curiosity module (Figure 4).

4.2. Utility of learned behaviors for downstream tasks

While the behaviors acquired are semantically meaningful, we would like to quantify their utility as pre-training for downstream tasks. We analyze the performance on `Lifting` where the agent must grasp and lift a block to a

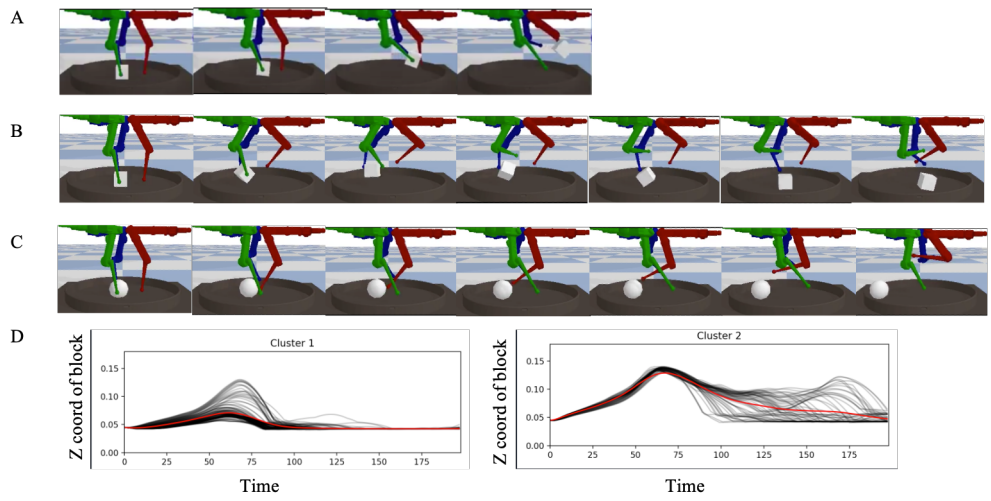


Figure 4. Examples of discovered behaviors. The agent discovers experimental behaviors that allow it to characterize each environmental object in a binary manner, e.g., heavy/light, big small, rollable/not rollable, etc. These behaviors are acquired without any external supervision by maximizing the causal curiosity reward. **A, B, C** correspond to self-discovered toss, lift-and-spin and roll behaviors respectively. **D** shows an ablation study, where the agent is only provided the z coordinate of the block in every environment. Each line corresponds to one environment and the z coordinate of the block is plotted with time when the discovered behavior is applied. It learns a lifting behavior, where cluster 1 represents the heavy blocks (z coordinate does not change much) and cluster 2 represents the light blocks (z increases as block is lifted and then falls when dropped and subsequently increases again when it bounces).

predetermined height and `Travel`, where the agent must impart a velocity to the block along a predetermined direction. We re-train the learnt planner using an external reward for these tasks (**Curious**). We implement a baseline vanilla Cross Entropy Method optimized Model Predictive Control Planner (De Boer et al., 2005) (**Vanilla CEM**) trained using the identical reward function and compare the rewards per trajectory during training. We also run a baseline (**Additive reward**) which explores whether the agent recieves both the causal curiosity reward and the external reward. We find high zero-shot generalizability and quicker convergence as compared to the vanilla CEM (Figure 5). We also find that additive rewards, achieves suboptimal performance due to competing objectives. For details, we refer the reader to the Supplementary Material D.

4.3. Visualization of hierarchical binary latent space

Our agents discover a disentangled latent space such that they are able to isolate the sources of causation of the variability they encounters in their environments. For every environment, they learn a disentangled embedding vector which describes each of the causal factors.

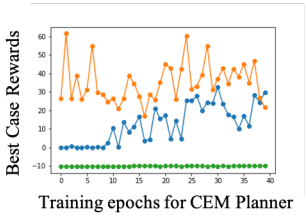
To show this, we use 3 separate experimental setups - `Mass`, `SizeMass` and `ShapeSizeMass` where each of the causal factors are allowed to vary over a range of discrete values. For details of the setup, we refer the reader to Supplementary Material B.2.

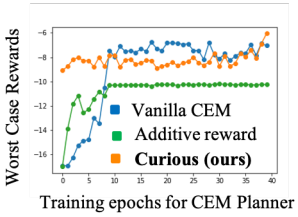
During training, the agent discovers a hierarchical binary latent space (Figure 3), where each level of hierarchy corresponds to a single causal factor. The binary values at each level of hierarchy correspond to the high/low values of the causal factor in question. To our knowledge, we obtain the first interpretable latent space describing the various causal processes in the environment of an agent. This implies that it learns to quantify each physical attribute of the blocks it encounters in a completely unsupervised manner.

4.4. Knowledge of causal factors aids transfer

Next, we test whether knowledge of the causal factors does indeed aid transfer and zero-shot generalizability. To this end, we supply the representations obtained by the agent during the experimental behavior phase as input to a policy network in addition to the state of the simulator, and train it for a place-and-orient downstream task (Figure 1). We define 2 experimental setups - `TransferMass` and `TransferSizeMass` where mass and size of the object in each environment is varied. In both setups, the agent learns about the varying causal mechanisms by optimizing causal curiosity. Subsequently, using the causal representation along with the state for each environment, it is trained to maximize external reward. For details of the setup, please see Supplementary Material C.

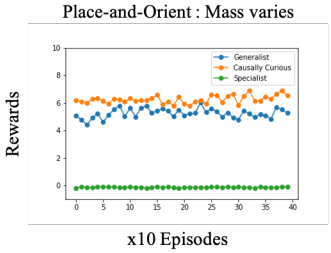
After training, the agents are exposed to a set of unseen test environments, where we analyze their zero-shot generalizability. These test environments consist of unseen masses





Discovered Experiments as Pre-training	
Method	**Time Steps of training**
Vanilla CEM Planner	$15\times10^5 \pm 122.47$
Curious (Ours)	6×10^5

Figure 5. Utility of discovered behaviors. We find that the behaviors discovered by the agents while optimizing causal curiosity show high zero-shot generalizability and converge to the same performance as conventional planners for downstream tasks. We also analyze the worst case performance and find that the pre-training ensures better performance than random initialization. The table compares the time-steps of training required on an average to acquire a skill with the time steps required to learn a similar behavior using external reward. We find that the unsupervised experimental behaviors are approximately 2.5 times more sample efficient. We also find that maxizing both curiosity and external reward in our experimental setups results in sub-optimal results.



Extra time steps needed by baseline to match CC performance	
Method	**Extra time steps of training**
Generalist - Mass	$4.8\times10^5 \pm 0.75\times10^5$
Generalist - Mass and Size	$9.8\times10^5 \pm 0.99\times10^5$

Figure 6. Knowledge of causal factors aids transfer. We find that knowledge of the causal representation allows agents to generalize to unseen environments with high zero-shot performance. The table depicts the extra timesteps required by the Generalist in each experimental setup to match the zero-shot performance of causally-curious agent. We find that as the number of varying causal factors increase, the difference in zero-shot performance of the Causally-curious agent and the Generalist increases, showing that the CC agents are indeed robust to multiple varying causal factors.

and sizes and their unseen combinations. This corresponds to "Strong Generalization" as defined by (Perez et al., 2020). We report results averaged over 10 random seeds.

For each setup, we train a PPO-optimized Actor-Critic Policy (referred to as **Causally-curious agent**) with access to the causal representations and an observation vector from the environment i.e., $\mathbf{a}_t \sim \pi(\cdot|\mathbf{o}_t, \mathbf{z}_{0:K})$. Similar to (Perez et al., 2020), we implement 2 baselines - the **Generalist** and the **Specialist**. The **Specialist** consists of an agent with identical architecture as **Causally-curious agent**, but without access to causal representations. It is initialized randomly and is trained only on the test environments, serving as a benchmark for complexity of the test tasks. It performs poorly, indicating that the test tasks are complex. The architecture of the **Generalist** is identical to the **Specialist**. Like the **Specialist**, the **Generalist** also does not have access to the causal representations, but is trained on the same set of training environments that the Causally-curious agent is trained on. The poor performance of the generalist indicates that the tasks distribution of training and test tasks differs significantly and that memorization of behaviors does not yield good transfer. We find that causally-curious agents significantly outperform the both baselines indicating that indeed, knowledge of the causal representation does aid zero-shot generalizability.

5. Conclusion

Our work introduces a causal viewpoint of POMDPs, where unobserved static state variables (i.e., causal factors) affect the transition of dynamic state variables. Causal curiosity rewards experimental behaviors an agent can conduct in an environment that underscore the effects of a subset of such global causal factors while obfuscating the effects of others. Motivated by the popular One-Factor-at-a-Time (OFAT) (Fisher, 1936; Hicks, 1964; Czitrom, 1999), our agents study the effects causal factors have on the dynamics of an environment through active experimentation and subsequently obtain a disentangled causal representation for causal factors of the environment. We discuss the implication of OFAT in Supplementary Material G. Finally, we show that knowledge of causal representations does indeed improve sample efficiency in transfer learning.

References

Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Wüthrich, M., Bengio, Y., Schölkopf, B., and Bauer, S. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-vae. arxiv 2018. *arXiv preprint arXiv:1804.03599*, 2018.

Camacho, E. F. and Alba, C. B. *Model predictive control*. Springer Science & Business Media, 2013.

Chaves, R., Luft, L., Maciel, T., Gross, D., Janzing, D., and Schölkopf, B. Inferring latent structures via information inequalities. In Zhang, N. L. and Tian, J. (eds.), *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pp. 112–121, Corvallis, OR, 2014. AUAI Press.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620, 2018.

Cuturi, M. and Blondel, M. Soft-dtw: a differentiable loss function for time-series. *arXiv preprint arXiv:1703.01541*, 2017.

Czitrom, V. One-factor-at-a-time versus designed experiments. *The American Statistician*, 53(2):126–131, 1999.

De Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

Doshi-Velez, F. and Konidaris, G. Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. In *IJCAI: proceedings of the conference*, volume 2016, pp. 1432. NIH Public Access, 2016.

Elwert, F. Graphical causal models. In *Handbook of causal analysis for social research*, pp. 245–273. Springer, 2013.

Fisher, R. A. Design of experiments. *Br Med J*, 1(3923): 554–554, 1936.

Grunwald, P. A tutorial introduction to the minimum description length principle. *arXiv preprint math/0406077*, 2004.

Hicks, C. R. Fundamental concepts in the design of experiments. 1964.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework.

Hill, A., Raffin, A., Ernestus, M., Gleave, A., Kanervisto, A., Traore, R., Dhariwal, P., Hesse, C., Klimov, O., Nichol, A., Plappert, M., Radford, A., Schulman, J., Sidor, S., and Wu, Y. Stable baselines. https://github.com/hill-a/stable-baselines, 2018.

Ilse, M., Tomczak, J. M., Louizos, C., and Welling, M. Diva: Domain invariant variational autoencoders. *arXiv preprint arXiv:1905.10427*, 2019.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Janzing, D., Chaves, R., and Schölkopf, B. Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New Journal of Physics*, 18(9):093052, 2016.

Killian, T. W., Daulton, S., Konidaris, G., and Doshi-Velez, F. Robust and efficient transfer learning with hidden parameter markov decision processes. In *Advances in neural information processing systems*, pp. 6250–6261, 2017.

Kim, H. and Mnih, A. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020.

Ngo, H., Luciw, M., Forster, A., and Schmidhuber, J. Learning skills from play: artificial curiosity on a katana robot arm. In *The 2012 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2012.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., and Schölkopf, B. Learning independent causal mechanisms.

In *International Conference on Machine Learning*, pp. 4036–4044. PMLR, 2018.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–17, 2017.

Pearl, J. *Causality*. Cambridge university press, 2009.

Perez, C. F., Such, F. P., and Karaletsos, T. Generalized hidden parameter mdps: Transferable model-based rl in a handful of trials. *AAAI Conference On Artifical Intelligence*, 2020.

Rissanen, J. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Schmidhuber, J. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pp. 1458–1463, 1991a.

Schmidhuber, J. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pp. 222–227, 1991b.

Schmidhuber, J. Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, 18(2):173–187, 2006.

Schmidhuber, J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

Schölkopf, B. Artificial intelligence: Learning to see and act (News & Views). *Nature*, 518(7540):486–487, 2015.

Schölkopf, B. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 1255–1262, 2012.

Spirtes, P. Introduction to causal inference. *Journal of Machine Learning Research*, 11(5), 2010.

Yao, J., Killian, T., Konidaris, G., and Doshi-Velez, F. Direct policy transfer via hidden parameter markov decision processes. In *LLARLA Workshop, FAIM*, volume 2018, 2018.

Zintgraf, L., Shiarlis, K., Igl, M., Schulze, S., Gal, Y., Hofmann, K., and Whiteson, S. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.

A. Important Note for Reviewers

The drive link to view visualizations is broken in the main text. To see videos please see footnote[1].

B. Implementation Details for Experiment Discovery

B.1. Planner

The Experiment Planner consisted of a uniform distribution planner for a horizon of 6 control signals. The planner was trained using the Cross Entropy Method Model Predictive Control (Camacho & Alba, 2013; De Boer et al., 2005) on the true environment. We sampled 30 plans per iteration from the distribution initialized to uniform $\mathcal{U}(controlLow, controlHigh)$. Each of the sampled plans are applied to each of the training environments and the top 10% of the plans are used to update the distribution. The CEM training required 10 iterations.

B.2. Training Environments

The training environments vary in each experiment. In Section 4.3, we utilize 3 setups, `Mass`, `SizeMass` and `ShapeSizeMass`. For `Mass`, we allow the agent to access 5 environments with masses varying from 0.1 kg to 0.5 kg. In `SizeMass`, the agent has access to 30 environments with masses varying uniformly from 0.1 to 0.5 kg and sizes from 0.05 to 0.1 meters. Finally, in `ShapeSizeMass`, the agent has access to 60 environments, with masses varying uniformly from 0.1 to 0.5 kg, sizes from 0.05 to 0.1 meters and shapes either being cubes or spheres. During experiment discovery, in each environment, the agent has access to the position of the block in the environment along with its quaternion orientation.

The total number of causal causal factors of each environment are rather large in number due to the fact that the simulator is a complex realistic physics engine. Examples of the causal factors in the environment include gravity, friction coefficients between all on interacting surfaces, shapes, sizes and masses of blocks, control signal frequencies of the environment. However, we only vary 1 during `Mass`, 2 during `SizeMass` and 3 during `ShapeSizeMass`.

B.3. Curiosity Reward Calculation

We predetermine the minimum description length of the clustering model $L(\mathbf{M})$ by assuming that the observations $\mathbf{o}_{0:T}$, obtained by applying experimental behavior $\mathbf{a}_{0:T}$ are produced by a bi-modal generator distribution, where each mode corresponds to either a low or high (quantized) value of a causal factor. This also ensures that $L(\mathbf{M})$ is as small as possible. The planner, eq. (7) solves the following optimization problem:

$$\arg\max_{\mathbf{a}_{0:T}\in\mathcal{A}^T}[\min\{d(\mathbf{o}_{0:T}, \mathbf{o'}_{0:T}) : \mathbf{o}_{0:T} \in C_1, \mathbf{o'}_{0:T} \in C_2\} - \max\{d(\mathbf{o}_{0:T}, \mathbf{o''}_{0:T}) : \mathbf{o''}_{0:T}, \mathbf{o}_{0:T} \in C_1\} - \max\{d(\mathbf{o'}_{0:T}, \mathbf{o'''}_{0:T}) : \mathbf{o'}_{0:T}, \mathbf{o'''}_{0:T} \in C_2\}] \tag{10}$$

the distance function $d(\cdot, \cdot)$ in the space of trajectories is set to be Soft Dynamic Time Warping (Cuturi & Blondel, 2017). The trajectory length T is 6 control steps long. The objective is a modified version of the Silhouette Score (Rousseeuw, 1987).

Intuitively, Objective (10) expresses the ability of a low complexity model, assumed to be bi-modal, to encode the state $\mathbf{O} = \mathbf{o}_{0:T}$. If multiple causal factors control $\mathbf{O}$, then the Minimum Description Length of $L(\mathbf{O})$ will be high. Subsequently, since $\mathbf{M}$ is a simple model, the deviation of $\mathbf{O}$ from $\mathbf{M}$ will be high i.e. $L(\mathbf{O}|\mathbf{M})$ will be high resulting in a low value of the optimization objective. C_1 and C_2 correspond to clusters of outcomes which quantize the values of a causal factor isolated by $\mathbf{a}_{0:T}$. $\mathbf{o}_{0:T}, \mathbf{o''}_{0:T} \in C_1$ correspond to trajectories of states i.e. observations obtained by applying $\mathbf{a}_{0:T}$ to environments with say, low values of a causal factor while $\mathbf{o'}_{0:T}, \mathbf{o'''}_{0:T} \in C_2$ correspond to trajectories of observations i.e. state obtained by applying $\mathbf{a}_{0:T}$ to environments with say, high values of the same causal factor. Objective (10) attempts to ensure that these clusters are far apart from each other and are tight i.e. a simple model $\mathbf{M}$ encodes $\mathbf{O}$ well.

C. Implementation Details for Transfer

In Section 4.4, we show the utility of learning causal representations in 2 separate experimental setups. During `TransferMass`, the agent has access to 10 environments during training, with masses ranging from 0.1 to 0.5 kg. At test time, the agent is required to perform the place-and-orient task masses 2 masses - 0.7 kg and 0.75 kg. During `TransferSizeMass`, the agent has access to 10 environments during training, with sizes from either 0.01 or 0.05 m and masses ranging from 0.1 to 0.5 kg. At test time the agent is asked to perform the task on 2 environments with masses 0.7 kg and 0.75 kg with sizes = 0.05 m.

We find that testing with large and light blocks increase the chances of accidental goal completions. Thus, during test-time, we use environments with high masses for out-of-distribution testing. The causal representation is concatenated to the state of the environment as a contextual input and supplied to a PPO-Optimized Actor-Critic Policy i.e., it receives 57 dimensional input for `TransferMass`,

[1] https://drive.google.com/drive/folders/13hZgzW_Tbd5EicLbIxUMQqD7KhfRSW9g?usp=sharing

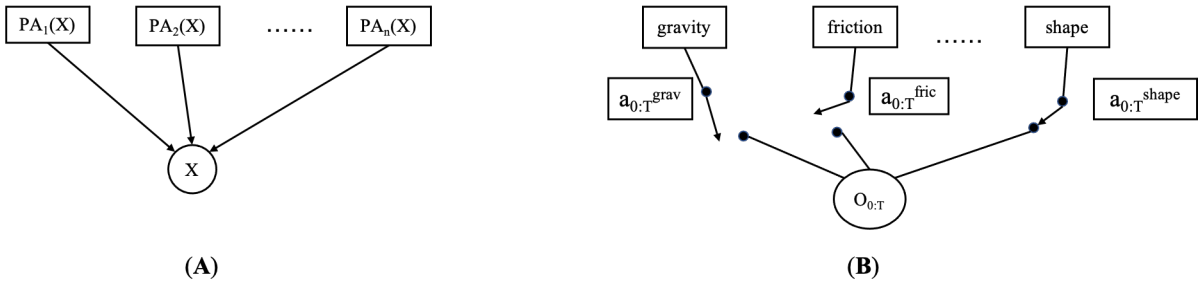


Figure 7. Directed Graphical Model. A Directed Acyclic Graph (DAG) visually represents the causal dependencies of observed and unobserved variables. In **(A)**, an observed variable $\mathbf{X}$ is caused by unobserved causal variables, $PA_i(\mathbf{X})$. In **(B)**, describes the scenario modeled in the paper, where a subset of the unobserved parent causal variables influence the observed variable $\mathbf{O}$. The action sequence $\mathbf{a}_{0:T}$ serves a gating mechanism, allowing or blocking particular edges of the causal graph.

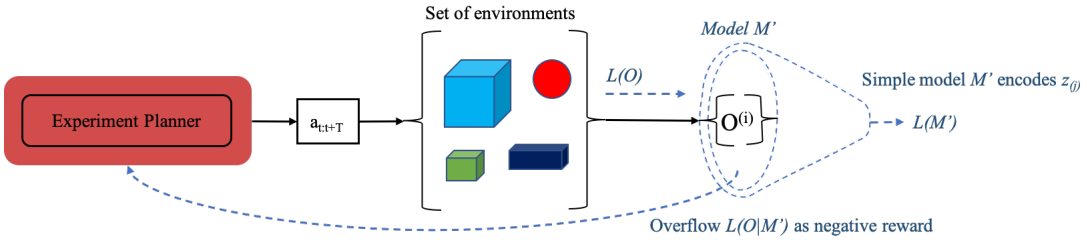


Figure 8. Overview of training. The experiment planner generates a trajectory of actions which is applied to each of the environments with varying causal factors namely mass, shape and size of blocks. For each environment, an observation trajectory or state $\mathbf{O}^i \in O$ is obtained. A simple model with fixed low expressive power is used to approximate the generative model for $\mathbf{O}$. The "information overflow" $L(\mathbf{O}|\mathbf{M})$ is returned as negative reward forcing O to be caused by few causal factors.

and a 58 dimensional for `TransferSizeMass`). The policy network consists of 2 hidden layers with 256 and 128 units respectively. The experiments are parallelized on 10 CPUs and implemented using stable baselines (Hill et al., 2018). The PPO configuration was {"gamma":0.9995, "n_steps": 5000, "ent_coef": 0, "learning_rate": 0.00025, "vf_coef": 0.5, "max_grad_norm": 10, "nminibatches": 1000, "noptepochs": 4}

The agent receives a dense reward at each time step during the maximizing external reward phase (Figure 1), the negative of the distance of the block from the goal position scaled by factor of 1000. The control signal was repeated 10 times to the actuators of the motors on each finger.

D. Implementation Details for Pre-trained Behaviors

In section 4.2, we study how the acquired experimental behaviors obtained through Causal Curiosity can be used as pre-training for a variety of downstream tasks. The Vanilla CEM depicts the cost of training an experiment planner from scratch to maximize an external dense reward where the agent minimizes the distance between the position of a block in an environment from the goal in the Lifting setup and imparts a velocity to the block along a particular direction in the Travel setup.

$$R(\mathbf{a}_{0:T}) = -\sum_t dist(\mathbf{goal}_t - \mathbf{block}_t) \tag{11}$$

The second baseline (Additive Reward) studies the setup when the agent receives both the curiosity signal and the external reward and attempts to maximize both. The agent receives access all the training environments with varying causal factors and must simultaneously maximize both curiosity and the task reward. The equation below shows the reward maximized for the `Lifting` task.

$$\begin{aligned} R(\mathbf{a}_{0:T}) = \sum_{envs}\sum_t^T -dist(\mathbf{goal}_t - \mathbf{block}_t) + \\ [\min\{d(\mathbf{o}_{0:T}, \mathbf{o'}_{0:T}) : \mathbf{o}_{0:T} \in C_1, \mathbf{o'}_{0:T} \in C_2\} - \\ \max\{d(\mathbf{o}_{0:T}, \mathbf{o''}_{0:T}) : \mathbf{o''}_{0:T}, \mathbf{o}_{0:T} \in C_1\} - \\ \max\{d(\mathbf{o'}_{0:T}, \mathbf{o'''}_{0:T}) : \mathbf{o'}_{0:T}, \mathbf{o'''}_{0:T} \in C_2\}] \end{aligned} \tag{12}$$

The curious agent first acquired the experimental behavior by interacting with multiple environments with varying causal factors. The lifting skill was obtained during `Mass`, when the agent attempted to differentiate between multiple blocks of varying mass. The curious agent trained for

$600,000$ time steps on the curiosity reward. The acquired behavior was then applied to the downstream lifting task and fine tuned to external rewards. The Vanilla CEM baseline had an identical structure to that of the Curious agent, and received only external reward as in Equation (11). The additive agent simultaneously optimized both external reward and the curiosity reward as in Equation (12).

We find that maximizing the curiosity reward in addition to simultaneously maximizing external rewards results in sub-optimal performance due to our formulation of the curiosity reward. To maximize curiosity, the agent must discover behaviors that divide environments into 2 clusters. Thus in the context of the experimental setups, this corresponds to acquiring a lifting/pushing behavior that allows the agent to lift/impart horizontal velocity to blocks in half of the environments, while not being able to do so in the remaining environments. However, the explicit external reward incentivizes the agent to lift/impart horizontal velocity blocks in all environments. Thus these competing objectives result in sub-par performance.

E. Intuition for Definition of Causal Factors

We begin with a simple example of a person walking on earth. This person experiences various physical processes while interacting in their world, for example gravity, friction, wind etc. These physical processes affect the outcome of interactions of the person with their environment. For example, while jumping on earth, the human experiences gravity which affects the outcome of their jump, the fact that they falls back to the ground. Additionally, these physical processes (or causal mechanisms) are parameterized by causal factors, for example, acceleration constant due to gravity $g = 9.8m/s^2$ on earth, or coefficients of friction between their feet and the ground which assume particular numerical values.

These causal factors may vary across multiple environments. For example, the person may walk on sand or on ice, surfaces with varying frictional values. Thus the outcome of running on such surfaces will vary, running on sand will require significant effort, while running on ice may result in the person slipping. Thus the coefficient of friction between the person's feet and the surface they walk on affects the outcome of a particular behavior in said environment. In our definition, $\mathbf{h}_j$ are causal factors such as friction or gravity etc. H is the global set containing all such causal factors.

Now we ask the question (which we subsequently answer), given multiple environments, how would a human characterize each of them depending on the value of a causal factor? Through experimental behaviors. The human in the above example would attempt to run in each of the environments she encountered, be it on sand, on ice, in mud etc.

Algorithm 2 Inference Loop

Input: Unseen Test Environment env, trained Planner and Causal Inference Module
Initialize $\mathbf{causalRep} = [\,]$
Initialize training environment set $Envs$
for k in range(K) **do**
  Reset env
  Sample experimental behavior $\mathbf{a}_{0:T} \sim$ CEM($\cdot|\,\mathbf{causalRep}$)
  Apply $\mathbf{a}_{0:T}$ to env
  Collect $\mathbf{O} = \mathbf{o}_{0:T}$
  Use learnt $q_M(\mathbf{z}|\mathbf{O}, \mathbf{causalRep})$ for cluster assignment i.e. $\mathbf{z}_k = q_M(\mathbf{z}|\mathbf{O}, causalRep)$
  Append z_k to $causalRep$
end for
Learn a policy conditioned on causal factors $\mathbf{a}_t \sim \pi(\cdot|\mathbf{o}_t, \mathbf{z}_{0:K})$ to maximize external reward.

If they slipped in an environment, she would characterize it as slippery. If they didn't, they would characterize it as non-slippery. We attempt to equip our agent with similar logic. The "sequence of actions" ($\mathbf{a}_{0:T}$) described in our paper corresponds to the human running. The sequence of observations ($\mathbf{o}_{0:T}$) corresponds to the outcome of running "experiment". $\mathbf{o}_{0:T}$ might belong to either of the clusters of outcomes C_a or C_b corresponding to slipping or not slipping.

F. Extensive Related Work

(Doshi-Velez & Konidaris, 2016) define a class Markov Decision Processes where transition probabilities $p(s_{t+1}|s_t, a_t; \theta)$ depend on a hidden parameter θ, whose value is not observed, but its effects are felt. (Killian et al., 2017) and (Yao et al., 2018) utilize these Hidden Parameter MDPs (Markov Decision Processes) to enable efficient policy transfer, assuming that transition probabilities across states are a function of hidden parameters. (Perez et al., 2020) relax this assumption, allowing both transition probabilities and reward functions to be functions of hidden parameters. (Zintgraf et al., 2019) approach the problem from a Bayes-optimal policy standpoint, defining transition probabilities and reward functions to be dependent on a hidden parameter characteristic of the MDP in consideration. We utilize this setup to define causal factors.

Substantial attempts have been made at unsupervised disentanglement, most notably, the β-VAE (Higgins et al.) (Burgess et al., 2018), where a combination of factored priors and the information bottleneck force disentangled representations. (Kim & Mnih, 2018) enforce explicit factorization of the prior without compromising on the mutual information between the data and latent variables,

a shortcoming of the β-VAE. (Chen et al., 2018) factor the KL divergence into a more explicit form, highlighting an improved objective function and a classifier-agnostic disentanglement metric. (Locatello et al., 2018) show theoretically that unsupervised disentanglement (in the absence of inductive biases) is impossible and highly unstable, susceptible to random seed values. They follow this up with (Locatello et al., 2020) where they show, both theoretically and experimentally, that pair-wise images provide sufficient inductive bias to disentangle causal factors of variation. However, these works have been applied to supervised learning problems whereas we attempt to disentangle the effects of hidden variables in dynamical environments, a relatively untouched question.

G. Scalability Limitation

We utilize the extremely popular One-Factor-at-a-time (OFAT) general paradigm of scientific investigation, as an inspiration for our method. In the case of many hundreds of causal factors, the complexity of this method will scale exponentially. However, we believe that this would indeed be the case given a human experimenter attempting to discover the causation in any system she is studying. Learning about causation is a computationally expensive affair. We point the reader towards a wealth of material on the design of scientific experiments and more specifically the lack of scalability of OFAT (Fisher, 1936; Hicks, 1964; Czitrom, 1999). Nevertheless, OFAT remains the de facto standard for scientific investigation.