

Parameter Estimation using Reinforcement Learning Causal Curiosity: Limits and Challenges

Miguel Arana-Catania^{1*} and Weisi Guo¹

¹Digital Scholarship at Oxford, University of Oxford, UK.

²Faculty of Engineering and Applied Sciences, Cranfield University, UK.

*Corresponding author(s). E-mail(s): humd0244@ox.ac.uk;

Abstract

Causal understanding is important in many disciplines of science and engineering, where we seek to understand how different factors in the system causally affect an experiment or situation and pave a pathway towards creating effective or optimising existing models. Examples of use cases are autonomous exploration and modelling of unknown environments or assessing key variables in optimising large complex systems. In this paper, we analyse a Reinforcement Learning approach called Causal Curiosity, which aims to estimate as accurately and efficiently as possible, without directly measuring them, the value of factors that causally determine the dynamics of a system. Whilst the idea presents a pathway forward, measurement accuracy is the foundation of methodology effectiveness. Focusing on the current causal curiosity’s robotic manipulator, we present for the first time a measurement accuracy analysis of the future potentials and current limitations of this technique and an analysis of its sensitivity and confounding factor disentanglement capability - crucial for causal analysis. As a result of our work, we promote proposals for an improved and efficient design of Causal Curiosity methods to be applied to real-world complex scenarios.

Keywords: reinforcement learning, dynamical systems, causal analysis, sensitivity analysis, parameter estimation

1 Introduction

In the study of dynamical systems, it is a common need to estimate the value of parameters that determine their interactions and dynamics. These may be unknown, for example, because they belong to an unexplored environment (e.g. the weight of a lunar rock to be picked up by a rover on a space mission, or the magnitude and direction of the wind field for drone operations), or because our sensors and system model presents errors or uncertainties in the value of its parameters (e.g. the distance between the wheels of a mobile autonomous robot may vary after interacting with

a rough terrain). In the case of autonomous systems, such as the examples above, it might not be feasible to access the system to perform measurements, or they might not have the appropriate tools because of design reasons. In these situations, it is crucial to have indirect strategies that allow estimating their value as accurately as possible. Here, we focus on a general and established experimental case of a robotic manipulator interacting with physical objects with unknown parameters, and the use of machine learning to estimate these parameters.

In this work, we approach this problem from the point of view of causal analysis. The enormous

development and application of machine learning techniques witnessed in recent years are allowing to discover some of their main limitations. One of these limitations is the fact that machine learning techniques are usually designed to find correlations in the data during the learning process but not to find causation or apply proper causal reasoning [1–7]. This results in models that are unable to correctly solve tasks that are nevertheless solvable with causal reasoning, e.g. avoiding spurious relationships due to data correlations or reasoning using the correct causal relations. Examples can be seen in fields such as computer vision [8, 9], reinforcement learning [10–12], natural language processing [13, 14], or graph analysis [15, 16]. As a result of this, an increasing number of works are trying to bridge this gap by combining causal analysis and machine learning.

From a causal point of view, the parameters to be estimated in our system are denoted as causal factors, since their values causally determine the dynamics of the system. The objective of the Causal Curiosity technique is to return strategies to interact with the unknown objects that are most effective in estimating these causal factors. In the language of causal analysis, we look for interventions that allow us to estimate the Causal Directed Acyclic Graphs (DAG) of our system [17–20]. In the case analysed in this work, this translates into identifying which are the most appropriate movements of our robotic manipulator to estimate the unknown parameters of the object (e.g., its mass or friction coefficient).

Bringing together the use of a machine learning technique in the estimation of different parameters and using a causal approach, we work here with the recent proposal of Causal Curiosity [21]. This approach has been designed to determine the fundamental parameters of an unknown system in a systematic and organised way, following a strategy of scientific experimentation on these factors. It shows an enormous potential since it proposes a robust, explainable methodology, anchored in a causal analysis framework that thus allows going beyond the previously mentioned limitation of common methodologies in Reinforcement Learning (RL).

This method presents fundamental advantages in comparison to other methodologies. Previous works on parameter estimation with robotic manipulators show the possibility of estimating,

for example, mass, stiffness, inertia, hardness, shape, texture, friction, etc [22–30]. However, these methodologies are not generalisable, in the sense that they can only estimate certain types of parameters and need specific formulations for each type of parameter (e.g. in general a method for estimating the inertia of an object cannot be used to estimate its friction coefficient). This is partly due to the fact that traditional methodologies commonly use knowledge of the equations that define the dynamics of the system (e.g. using a linear approximation of the equations to obtain the relationship between parameters). In our case, it is sufficient to simulate the dynamics of the system, without the need for explicit knowledge of the system equations or to perform operations on them to understand the effect of the parameters to be estimated on the observations of the system. Thus, the method is generalisable, in the way that it can be applied to any type of system parameter, and do so without main methodological changes in the implementation regardless of the parameter we consider. A second fundamental advantage is the ability to obtain an explainable methodology. As we will see later, the methodology maximises a human-explainable reward. Furthermore, the formulation of the problem is done in a causal framework, where the parameters affecting the system and the causal relationship between them are clearly defined, which is fundamental for its explainability. As a third fundamental advantage, the causal approach is an upgrade over non-causal machine learning methodologies. The latter do not consider in any way the causal relationship of the parameters on which they act and therefore fail in the case of confounding relationships between the variables on which they are trained. In our case, the analysis of the system always takes into account the causal relationships and therefore identifies beforehand the possible limitations that may exist, and the impact that these may have on the estimation of the parameters, as we will see especially in the last experiments carried out in this work.

Although there are many advantages to this methodology, the original proposal of this approach [21] leaves unanswered how much of this potential can be realised in practice when dealing with complex systems.

The main contribution of this work is the analysis of this technique, in this case examined in the

use case of a robotic manipulator, by confronting it for the first time with limiting situations in the complexity of the system to be explored or in the sensitivity sought in the parameter estimation, with the aim of understanding its limits and challenges. As an additional contribution, we also propose changes in the methodology to increase its effectiveness.

The implementation of the robotic manipulator is done using the same simulation framework as the article of the original proposal, CausalWorld¹, which is a framework specifically designed to work with causal approaches.

Our results confirm the enormous potential of the methodology and point out scenarios where difficulties are encountered and new lines of future work are needed to advance this methodology. Among other results, regarding its potential, we show the possibility of refining the accuracy of parameter identification by several orders of magnitude while increasing the robustness of the analysis by about 15% in the case of our proposed methodology. In relation to its limitations, we show the impossibility of solving some complex experimental situations studied for the first time here with multiple parameters varying simultaneously. We also analyse, with positive results, cases in which the factors present causal relationships not only with respect to the general evolution of the system dynamics, but also between them, and in particular we study a case with confounding variables.

We have divided our analysis into five strands, in order to obtain a result as comprehensive as possible. We focus our research questions (RQ) on developing the measurement accuracy analysis of the following areas that enable improved causal analysis:

- RQ1. Accuracy in the estimation of different causal factors.
- RQ2. Granularity in the determination of causal factors.
- RQ3. Gap size effect in the estimation of causal factors.
- RQ4. Multiple causal factors determination.
- RQ5. Causally related and confounding causal factors estimation.

Additionally, we evaluate the use of Proximal Policy Optimization (PPO) [31] as a causal factor estimation method and compare it to the Cross-Entropy Method (CEM) optimised Model Predictive Control Planner [32, 33] proposed by the authors of Causal Curiosity.

2 Related Work

The fields of Causal Analysis [17–20] and Reinforcement Learning are increasingly being combined as a way of overcoming the inevitable limitations of the traditional correlational approach to RL that give rise to problems of generalisability, robustness and inefficiency, among others. This is leading to the development of the emerging fields of Causal Machine Learning and Causal RL [1, 2, 34–41].

In this latter field, several works introduce causal concepts to try to improve the inefficiency in exploring the state space, one of the main problems in RL (traditionally not related to causality, see Amin et al. [42]). For example, Peng et al. [43] propose a causality-driven hierarchical RL framework using causality-driven exploration and subgoal structures to improve the exploration phase in challenging tasks with sparse rewards, Rezende et al. [44] work with partial models as a way to avoid learning the full model in intractable high-dimensional spaces, and use causality to correctly connect such partial models, Pitís et al. [45] introduce local causal models, which are induced from a global causal model by conditioning on a subset of the state space, to improve the sample efficiency and performance of RL systems that involve sub-processes, Seitzer et al. [46] improve exploration by using the local causal graph to guide the learning based on what can be influenced in each state, Molina et al. [47] use knowledge from causal models to reduce the exploration space of RL models.

A key element in our approach is the concept of curiosity. The field of RL has long considered intrinsic motivation approaches and curiosity-driven learning as a way to solve complex tasks and deal with the exploration-exploitation dilemma [48–68].

However, Causal Curiosity is not only linked to efficiently exploring the state space as in most of the previous works but it is ultimately related

¹<https://github.com/rr-learning/CausalWorld>

to the goal of correctly identifying causal factors through experimentation. Other works with a focus on intervention and experimentation to discover causal relationships are the following: Gasse et al. [69] propose a model-based Causal RL method where the information obtained from the interactions of the agent with the environment are combined with observational offline data from another agents interactions, Dasgupta et al. [41] use meta-learning to train agents in tasks depending on a causal structure, and show how the agents become capable of performing experiments that can be seen as causal reasoning, Ke et al. [70] study causal induction in model-based reinforcement learning, Nair et al. [71] propose techniques for causal induction from raw visual observations and causal graph encoding, Ding et al. [72] propose a method that alternates between performing interventions to estimate the causal graph and using the graph to learn generalisable models, Thomas et al. [73] work on objective functions to obtain a disentangled representation through interaction with the environment, [74–76] work in obtaining disentangled representations using variational autoencoders, Volodin et al. [77] intervene in an unknown environment to solve spurious correlations.

3 Methodology

In this section, we introduce the conceptual framework of causal analysis used in this paper [17–20] and summarise the main theoretical elements of the Causal Curiosity approach [21] and the details of the specific methodology applied in this paper.

3.1 Simulation framework and main parameters

The framework used in this research is CausalWorld². This is an open-source simulation framework and benchmark for causal structure and transfer learning in a robotic manipulation environment. This framework has been designed to allow an easy analysis of a dynamic system from a causal point of view, enabling in addition the application of reinforcement learning methodology.

CausalWorld uses the Bullet physics engine [78] to simulate the open-source TriFinger robot platform [79]. This framework allows the modification of different physical parameters that causally determine the evolution of the system’s interaction and dynamics. A screenshot of the simulation can be seen next in Figure 1 and videos of the robot in operation can be found in the framework site².

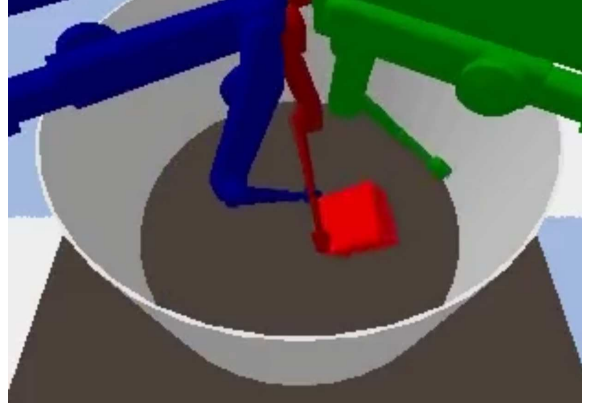


Fig. 1: CausalWorld simulation

In this work, the robot manipulator interacts with a single object, a cube whose dynamic is defined by the following causal factors: **mass**, **size**, **lateral friction**, **spinning friction**, and **gravity**. The first four factors affect the object, while the last one affects both the object and the robot.

In Table 1 we present the initial value of these parameters and the maximum range of variation that we explore.

Factor	Value	Range
Mass	0.25	[0.01,0.5]
Size	0.075	[0.05,0.1]
Lat. Frict.	1	[0.1,1.0]
Spin. Frict.	0.001	[0.001,1.0]
Gravity	-9.81	[-1.0,-11.5]

Table 1: Causal factors values and ranges.

3.2 Causal framework

From the point of view of causal analysis [17–20], the variables x_i that define the system are related

²<https://github.com/rr-learning/CausalWorld>

to each other by a series of functions f_i called structural assignments:

$$x_i \equiv f_i(\epsilon_i; pa_i) \quad (1)$$

These functions determine how each variable changes influenced by the other variables (called parents, and denoted by pa_i), and by independent exogenous noise variables ϵ_i .

The set of structural assignments is called the Structural Causal Model (SCM):

$$S = \{f_i\}_{i=1}^N \quad (2)$$

The SCM is a complete description of the system from a causal point of view and entails its observational, interventional and counterfactual distributions.

As will be seen in more detail below, in our case these structural assignments are the dynamical equations of the system that connect the parameters defining the object with which we will interact (mass, size, etc.) with the observations during the experiments (represented as the trajectories of the object).

Commonly, several principles are generally assumed [19]. Firstly, the Principle of Independent Mechanisms, which from a probabilistic point of view indicates that the conditional distribution of each variable given its causes (i.e. its mechanism) does not inform or influence the other conditional distributions.

This principle allows us to disentangle the chain rule of probability,

$$p(x_1, \dots, x_n) = \prod_i p(x_i | x_1, \dots, x_{i-1}) \quad (3)$$

and express the probability of the system as the product of probabilities of variables that depend only on their parents through the Markov factorisation property:

$$p(x_1, \dots, x_n) = \prod_i p(x_i | pa_i) \quad (4)$$

We also assume that the variables are related in a non-cyclic manner which implies that every SCM induces a Causal Directed Acyclic Graph (DAG) as depicted in Figure 2. In this figure, we represent the DAG corresponding to the system studied in this article, with the Observation node representing the trajectory of the object, which

is affected, as indicated by the arrows, by the different parameters mentioned above.

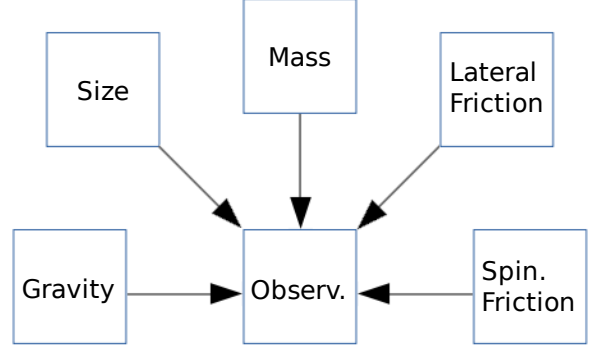


Fig. 2: Causal DAG

3.3 System dynamics, causal factors and reinforcement learning

In addition to the SCM representing the causal relationships of the system, from a dynamic point of view the system is defined by a Partially Observable Markov Decision Process (POMDP) [80]. This is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, E, R)$, where \mathcal{S} is the state space, \mathcal{A} is the action space, \mathcal{O} is the observation space, P describes the dynamics of the system through a conditional probability distribution $P(s_{t+1} | s_t, a_t)$ with $s \in \mathcal{S}$ and $a \in \mathcal{A}$, E is the emission function defining the distribution $E(o_t | s_t)$ with $o \in \mathcal{O}$, and R is the reward function. In this case, the states and observations are determined by the robot agent's actions and the previous causal factors (mass, size, etc.). The **observations** represent the object trajectory, and the **actions** the movements of the robot manipulator.

Each simulation is defined by an environment, for which we can define variations in which a subset of factors modify its value (e.g. we will consider various values for the mass of the object). In causal language, we represent setting each parameter to a specific value by using the $do(\cdot)$ operator, which indicates an intervention on that parameter. We recall here the difference between an observational probability distribution of an o variable $P(o | d')$ in which we observe the distribution of o given that we observe the variable d for a certain fixed value d' (i.e. we limit our observations to only

those with $\mathbf{d} = \mathbf{d}'$) and the interventional probability distribution $P(\mathbf{o} \mid do(\mathbf{d}'))$ representing the distribution of \mathbf{o} if we set $\mathbf{d} = \mathbf{d}'$. These two distributions are not equivalent.

The causal factors \mathbf{h} that we want to estimate are parameters such that by applying a certain sequence of actions on the different variations of an environment, the probability of the observations vary in such a way that the trajectories of the objects are organised in distinguishable disjoint sets according to the values of the causal factors:

$$p(\mathbf{o} \mid do(\mathbf{h} = \mathbf{h}'), \mathbf{a}) \neq p(\mathbf{o} \mid do(\mathbf{h} = \mathbf{h}''), \mathbf{a}) \quad (5)$$

E.g., if we set the mass of the object to values with very low and very high mass, we expect these two groups of environments to produce object trajectories that can be organised in two disjoint sets such as large and small displacement trajectories.

The sequence of actions that cluster the trajectories in a distinguishable way is not known in advance, and thus we need strategies to search for these sequences.

A representation of the RL interaction can be seen in Figure 3.

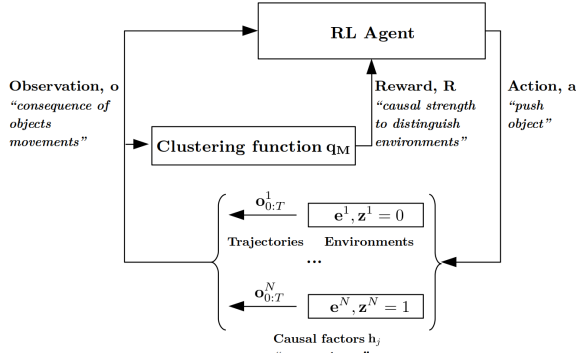


Fig. 3: RL interaction

To clarify all this, we explain next in more detail the case analysed in this work. In each one of our experiments, we consider a set of N environments \mathbf{e}^i , each of them determined by a set of K causal factors \mathbf{h}_j . In each interaction, we determine the value of only one of these causal factors, which presents a bimodal distribution. The interaction of the robot with the object aims to produce observations in the object's trajectory

$\mathbf{o}_{0:T}^i$ for T timesteps that can be identified with two clusters corresponding to the two modes of the causal factor value. For example, the objects may have values for the mass belonging to two classes: light objects and heavy objects; a certain thrust produces two distinguishable behaviours (whose trajectories can therefore be clustered into two groups): in the light case the object falls on one of its sides, and in the heavy case the object oscillates and return to its initial position. The method explores different movements of the robot until it finds that particular thrust that is able to clearly distinguish between the two groups of environments.

In all experiments, all ranges of values are partitioned in such a way that the two ranges have the same size which is also equal to the difference between the two ranges. The only exception is presented in Section 4.3 where this alternative is explicitly investigated.

In the initial Causal Curiosity proposal, all other causal factors were held constant in each individual experiment. In our case, in several experiments (see sections 4.4 and 4.5) other factors vary simultaneously, to study how the combined effect of these variations affects the estimation of the main causal factor.

In our experiments, we investigate in a practical way the limits and challenges in finding the appropriate sequences of actions that distinguish between the different factors (i.e. the actions that estimate the parameters) and how distinguishable these factors are (i.e. what is the accuracy of the estimation).

3.4 Parameter distinguishability and rewards

The next element to explain in our methodology is how we quantify whether the trajectories are distinguishable in terms of the causal factors. For this, we need to define a reward to quantify it, which will guide the search for optimal actions.

In the initial proposal, it was chosen to minimise the Minimum Description Length as a substitute for the Kolmogorov Complexity [81, 82]. In practice, and considering the bimodal character of the causal factor, this was implemented as a trajectory clustering problem using a Silhouette Score S [83] with a distance in the trajectory space defined by Soft Dynamic Time Warping [84].

In our case, we keep this as part of our reward. However, since we are exploring the limits of this methodology, we face environments where it is not possible to correctly identify all the environments. This implies the need to introduce an additional value that quantifies the correct identification of the scenarios. We implement this value by means of an F1 classification score C . Hence, our reward is given by the sum of these two elements:

$$R = C(q_M(\mathbf{o}_{0:T}^i), \mathbf{z}^i) + k \cdot S(\mathbf{o}_{0:T}^i) \quad (6)$$

where q_M is the clustering function learnt from the trajectories, $\mathbf{z}^i \in \{0,1\}$ is the true cluster membership identifying the two groups of environments, and k is a parameter defined to weight the importance of the two reward objectives, in our case set to 0.1 since the correct identification of the factors is more relevant than having better-defined clusters.

To summarise, our reward aims to correctly identify each trajectory as belonging to the correct type of environment (e.g. the set of trajectories belonging to high-mass objects) as well as make the two sets of trajectories as distinguishable as possible.

An example may clarify how the F1 value and the clustering score are obtained in our experiments. We set up two sets of environments. In the first set, all objects have a low value of the mass, while in the second set, they have a high value. Each of these two values is defined as a range so that each object has a different value of the mass within the range. We then perform a robot movement action (performed in several timesteps), which will result in a displacement of the object. The trajectories are used to define a clustering function that identifies the two sets of objects. Finally, we perform the same action on new objects and use the obtained trajectories and the clustering function to identify them as belonging to the set of low or high masses. If we evaluate the results of these, we obtain on the one hand the F1 classification value (related to the number of correct classifications) and on the other hand the clustering score (indicating the distinguishability of the trajectories). In the next section, we will explain the optimisation process of the robot's actions, in order to maximise these scores. As we will see from the results of the experiments, in

most of them it is possible to correctly classify objects with a score value of 1.0. In relation to this score, it is important to emphasise that the only goal is to assign one of the two possible categories correctly, so it is not surprising that this is not a problem. In fact, the original paper on this proposal does not include this score, probably for this reason. However, as we will see later, one of the aims of this project is to test the limits of the methodology by subjecting it to increasingly complex scenarios. In the last experiments, we will see how this classification is no longer correct in different scenarios. On the other hand, in all the experiments we also offer the value of the clustering score, which is never perfect, and which will serve to evaluate and compare in more detail all the results.

The full reward will be used both to drive the training of our RL system and to evaluate the performance of the results. In this respect, we will add details about this choice. The first of the elements considered, the Silhouette Score, is a common choice in the evaluation of the distinguishability of sets [85–87]. Additionally, it is the choice proposed in the original formulation of the method used, so it is useful for us to present its value in our experiments to compare our results with what would have been obtained in the original proposal. The second element, introduced as a novelty with respect to the original proposal, is the use of a score to evaluate the classification aspect. This is a fundamental point of our proposal since as we will see in the experiments carried out, especially in the last ones, when we confront this methodology with highly complex scenarios, it fails to correctly classify all the environments. This did not occur in the initial proposal, due to the simplicity of the scenarios analysed, which is similar to our first experiments, but it is an important element in this analysis of the limits and challenges of the methodology. Multiple standard metrics can be considered as classification scores, such as precision, recall, accuracy or F1. Our choice of F1 as the metric used represents an appropriate compromise, as it is the harmonic mean between precision and recall, which represents a good balance between them.

3.5 Optimisation of interactions

The last main element of the methodology is the definition of the approaches used to perform the search for optimal actions maximising the above reward. Here there is also a relevant difference in our methodology compared with the initial Causal Curiosity proposal. In the initial proposal, the Cross-Entropy Method (CEM) optimised Model Predictive Control Planner [32, 33] is used. It works as follows: In each iteration, the planner proposes a set of plans sampled from a uniform distribution of the robot movements. Each plan is a sequence of movements, generated by applying control signals to the actuators in the joints of the robot. Using a horizon equal to 6 implies that during the experiment each joint receives 6 control signals, evenly spaced in time. A set of plans is proposed and each of the plans is executed in each of the environments. For each plan, we obtain a set of observations, determined by the positions of the object we interact with. The observations of each plan in the set of environments are used to obtain the reward for that plan. The subset of plans with the highest reward is selected to update the initial distribution from which we sampled the plans. This distribution is adjusted to fit the distribution related to those most successful plans. Once this is done, the process is repeated for another iteration. In this new iteration, the previously successful plans are more likely to be sampled. This will lead to the convergence of the plans that maximise the reward. As can be seen in this methodology, this is an open-loop control system, the actions are generated at the beginning of each experiment and thus are decoupled from the observations happening during the interaction.

This method is valid for simple scenarios but may be insufficient in more complex scenarios due to the lack of connection between action planning and observations. Therefore, in addition to this method, in this work we apply a PPO planner. This new planner has two fundamental differences with respect to the previous one. Firstly, we use a neural network to generate the actions of the robot. Secondly, this policy network is modified using a PPO method [31]. The process is carried out in the following steps: The policy network generates actions for the robot that similarly to the previous case are executed in the set of environments. As in the previous case, actions are

generated as control signals applied regularly to the actuators in the joints of the robot. This produces a change of state in the environments with their corresponding observations, which together with their reward are used as input to produce an update in the policy network. Once this is done, a new set of actions is generated and the cycle is repeated. The basis of the implementation used has been provided by Stable Baselines³, on which we have made a number of modifications to implement the previous process. The execution of PPO is performed using multiprocessing vectorised environments. Each of the environments running in parallel corresponds to an environment e^i . Using wrappers we modified the actions so that all the environments reproduce the same actions (which by default does not happen since in general, the objective of parallelisation is to probe different actions of the same policy), and we introduced a shared value of the reward, calculated in the last timestep using the observations of all the environments (by default each environment only uses its own observations to calculate the reward, which in this case is insufficient since the reward is defined by comparison between environments). These modifications allow us to use a more efficient method such as PPO, while maintaining a similar approach to the original method. In the original article, PPO is only used to carry out tasks subsequent to the estimation of the causal factors, while in our case this method is used directly for the estimation of the factors themselves. Throughout this paper, we will present in parallel the results using CEM and the results using PPO. The former corresponds to the methodology used in the original article, and the latter in this case. In this way, we will be able to compare the differences between both methodologies for each experiment.

3.6 Implementation and experimental details

Regarding the implementation details, the CEM planner has been implemented from the code of the Causal Curiosity repository⁴. The code has been optimised to allow multiprocessing and to allow

³<https://stable-baselines3.readthedocs.io>

⁴<https://github.com/sumedh7/CausalCuriosity>

the simultaneous variation of several causal factors. The output has also been modified to obtain the new reward proposed by us in this section.

The experiments are carried out as follows. For each experiment, 20 environments are created, half of which corresponds to one of the ranges of the parameter to be estimated and half to the other range (e.g. light masses and heavy masses). An action is then executed in the environment in 6 movements. This action is executed in the 20 environments, obtaining 20 trajectories of the objects. The trajectories are used to define a clustering function, which allows to obtain the scores measuring the distinguishability of the parameter sets.

In each experiment, there are 100 replications of this process. As will be seen from the results presented in the next section, in several experiments the value of the scores is low. In the case of using this procedure in a real use case, the process could be optimised to further improve the score obtained. In our work, the aim is to show the limitations of the methodology in a comparative way, so we carried out all the experiments at the same optimisation levels, to allow the differences between the experiments to be shown, in order to understand the limits of the experiments.

The experiments can be reproduced by following these steps using the code provided in the original paper⁴. It does not provide the value of the classification score, but this can be obtained directly from the trajectories by comparing them with the clustering function.

The hyperparameters relevant to this methodology are the following: environments = 20, plans per iteration = 5, iterations = 100, ratio of plans selected between updates = 0.4, frames per episode = 198, plan horizon = 6. The PPO planer has been implemented from the above code and the PPO2 implementation of Stable Baselines, with the modifications mentioned above. The relevant hyperparameters are: environments = 20, discount factor gamma = 0.9995, entropy coefficient ent_coef = 0, learning rate = 0.00005, value function coefficient vf_coef = 0.5, maximum value for the gradient clipping max_grad_norm = 10, number of epochs when optimising the surrogate noptepochs = 4, iterations = 500, MlpPolicy network with 2 hidden layers with sizes 256 and 128, frames per episode = 198, plan horizon = 6.

These hyperparameters have been defined without the need for hypertuning processes because as mentioned above, the aim of the experiments is not to obtain the best possible result tuned to each scenario but to provide a comparative result between the different scenarios, in order to show the scenarios that offer the greatest challenges to the methodology. However, in real use cases, it may be relevant to set the parameters for better performance. For each dynamic system, a different set of parameters may be the most relevant in their effect on performance. Multiple studies analyse the effect of these parameters and propose ideas for optimal selection [88–90]. We recommend the use of libraries to optimise the search for suitable parameters. Optuna⁵, Hyperopt⁶, and Ray-Tune⁷ are commonly used libraries for this purpose. They allow to parallelise searches, which significantly increases their efficiency. Another advantage they offer is the possibility of using more advanced search techniques such as Bayesian optimisation, in which the degree of uncertainty over the parameter space is quantified and used to select new regions to explore. This methodology is much more effective than a blind or a grid search.

The rationale for the selection of parameters to be estimated has been to offer the largest possible set of parameters that define the dynamics of the implemented case. The platform used for the simulations restricts the definition of solid objects⁸ to the parameters of mass, size, lateral friction and spinning friction, and the gravity parameter affecting the dynamics of the interaction between the robot and the object. As we will see in the results, especially in the more complex experiments, each type of parameter affects differently the dynamics of the experiments and thus the effectiveness of the methodology.

All the experiments will follow this methodology. In the next section, we describe the details of each of the experiments.

⁵<https://optuna.org/>

⁶<https://hyperopt.github.io/hyperopt/>

⁷<https://docs.ray.io/en/latest/tune/index.html>

⁸<https://causal-world.readthedocs.io/en/latest/modules/envs/envs.html#rigidob>

4 Experiments

The experiments carried out correspond to each of the five research questions presented in Section 1.

In the following, we present each experiment and its results.

4.1 Comparison of estimation of different causal factors

The first experiment aims to provide a basis for comparison of subsequent more complex scenarios. Here we analyse the efficiency of the methods in correctly identifying causal factors using a wide range of values, and examine whether there are notable differences according to the type of causal factor at this level of experimentation.

In Figure 4 we can see a schematic representation of the values of one of these experiments, with the mass values organised into two groups represented by the two horizontal lines above. As can be seen in the figure, in this case, there is a set of objects with masses between 0.01 and 0.17, and another set of objects with masses between 0.33 and 0.5. We expect to find the sequence of interactions with the objects, which given a new object of unknown mass, maximises the correct identification of its mass (as belonging to one or the other range of masses).



Fig. 4: Mass values

In Table 2 we show the results of the experiment. We present the two scores independently: F1 classification and clustering score, and recall that the former indicates that the identification of the factors is properly performed, while the latter represents how distinguishable the clusters are from each other. As can be seen, this experiment does not present any identification problem, since all the F1 scores are 1.0, which means that all the trajectories are properly identified as belonging to one set of masses or the other. The results show similar clustering scores regardless of the factor under investigation. This shows that at this level the robustness in the estimation of all of them is

similar, none of them seems to be more challenging in this initial experiment. As a comparison between the two optimisation strategies, if we take the mean value of the clustering scores, we obtain a value of 0.896 ± 0.038 for the CEM planner and 0.873 ± 0.045 for the PPO planner. However, we should remember that they represent qualitatively different parameters. This experiment provides some initial numbers to evaluate the accuracy of the estimation of different causal factors (RQ1) in a simple scenario and will be used for comparison in the following experiments when analysing more complex scenarios.

4.2 Granularity precision in causal factor determination

In this experiment, we analyse a more complex scenario, where our aim is to assign a more precise value to the causal factor and not just categorise it into two ranges. The objective is to understand how being more precise in the determination of the values (i.e. reducing the granularity of the considered ranges) affects the results (RQ2). Following the initial approach, what we do is take the two initial clusters and further split them into two other clusters. This procedure is repeated up to six times. In Figure 5 we see the representation of this procedure, and how the last clusters present a high precision in the knowledge of the value compared to the first ones (in the figure we represent only the first three bipartitions, instead of the six that we have performed). We note that although we follow a simple bipartition model here, the clusters can be organised with a uniform spacing at the desired level of precision, to better represent the parameter range. To clarify the level of precision we are talking about, consider that the ratio between the length of the range of values explored in the first experiment and this one is 0.0014. That is, if the strategy is successful, we are able to estimate each parameter with an accuracy three orders of magnitude higher.

Table 3 shows the results of this experiment. To synthesise the results, we present here values at the last level, and thus with the highest precision in identifying the factors. At this level, and according to the previous figure, there are 64 pairs of clusters. We present 2 pairs of clusters in the table, corresponding to the left beginning and right end

Causal factor		Range	CEM planner		PPO planner	
			F1	Clust.	F1	Clust.
Mass	F	[0.01,0.1733][0.3367,0.5]	1.00	0.909	1.00	0.836
Size	F	[0.05,0.0667][0.0833,0.1]	1.00	0.879	1.00	0.815
Lateral Friction	F	[0.1,0.4][0.7,1.0]	1.00	0.949	1.00	0.889
Spinning Friction	F	[0.001,0.334][0.667,1.0]	1.00	0.846	1.00	0.914
Gravity	F	[-1.0,-4.5][-8.0,-11.5]	1.00	0.896	1.00	0.912

Table 2: Comparison of estimation of different causal factors



Fig. 5: Successive bipartition of causal factor ranges. The last three levels are omitted

of the range, and identified in the table respectively with the initials L and R. In this way, we can also take into account the effect of the factor being at one end or the other of the range (i.e. discerning between two light masses may be different to discerning between two heavy masses). To compare the change in the effectiveness of the experiment, we also include in the table the results of the first experiment over the full range, identified as F.

From the results, we can see that even at this high level of precision it is possible to correctly estimate all causal factors. Additionally, we see in the value of the clustering score that in general the clusters are more defined than in the first experiment. It is important to note that this is not an effect derived from the score itself, since what the score measures, the relationship between the size of each cluster and the spacing between them, remains constant at all levels. What this result indicates is that the robot is able to find more efficient strategies when dealing with quasi-point distributions, than in the case of distributions with larger spreads, which is a positive outcome since this may be often the use case of interest (it is in general more useful to estimate parameters with high precision, and not just broad ranges). As a general evaluation, if we compare the clustering scores of the first and last level, they increase on average by 5.6% and 6.7% for the L and R cases respectively using the CEM planner. For the

PPO planner case, the improvement is even much greater, with 12.0% and 14.4% for L and R respectively. We observe now a difference between the two methods used, CEM and PPO planners. In the previous experiment, the results of both methods were essentially equivalent. Now, when confronted with a more challenging experiment, we see that PPO shows better results in the clustering of the factors.

To better clarify the level of range accuracy achieved in this experiment, we can take the average value of each range as representative of the same: for example in the case of Spinning Friction, at the first level (F) we are discerning between a value of 0.17 and 0.83, while at the last level (L) we are able to distinguish between a friction value of 0.0012 and a value of 0.0021.

4.3 Gap size effect in the determination of different factors

In this case, the objective is to investigate how well-defined the ranges can be. Here we examine the effect in the results of reducing the gap between clusters in relation to their size (RQ3). Following the idea of the previous experiments, we start from an initial gap with the same size as each cluster and in each new experiment we divide this gap by half. We repeat this up to four times, going from a gap to cluster size ratio of 100% to 4.2% in the last case. Figure 6 presents these scenarios.

The results of the experiment are presented in Table 4. We include the case denoted as G with the shortest gap after 4 partitions and again the initial reference scenario denoted as F.

As can be seen from the results, in this case, we encounter some difficulties. In the case of two

Causal factor		Range	CEM planner		PPO planner	
			F1	Clust.	F1	Clust.
Mass	F	[0.01,0.1733][0.3367,0.5]	1.00	0.909	1.00	0.836
	L	[0.01,0.0102][0.0104,0.0107]	1.00	0.910	1.00	0.964
	R	[0.4993,0.4996][0.4998,0.5]	1.00	0.989	1.00	0.997
Size	F	[0.05,0.0667][0.0833,0.1]	1.00	0.879	1.00	0.815
	L	[0.05,0.05002][0.05005,0.05007]	1.00	0.987	1.00	0.992
	R	[0.09993,0.09995][0.09998,0.1]	1.00	0.996	1.00	0.991
Lateral Friction	F	[0.1,0.4][0.7,1.0]	1.00	0.949	1.00	0.889
	L	[0.1,0.1004][0.1008,0.1012]	1.00	0.993	1.00	0.994
	R	[0.9988,0.9992][0.9996,1.0]	1.00	0.998	1.00	0.997
Spinning Friction	F	[0.001,0.334][0.667,1.0]	1.00	0.846	1.00	0.914
	L	[0.001,0.0015][0.0019,0.0024]	1.00	0.955	1.00	0.993
	R	[0.9986,0.9991][0.9995,1.0]	1.00	0.803	1.00	0.999
Gravity	F	[-1.0,-4.5][-8.0,-11.5]	1.00	0.896	1.00	0.912
	L	[-1.0,-1.0048][-1.0096,-1.0144]	1.00	0.879	1.00	0.933
	R	[-11.4856,-11.4904][-11.4952,-11.5]	1.00	0.998	1.00	0.999

Table 3: Analysis of granularity in the estimation of causal factors

Causal factor		Range	CEM planner		PPO planner	
			F1	Clust.	F1	Clust.
Mass	F	[0.01,0.1733][0.3367,0.5]	1.00	0.909	1.00	0.836
	G	[0.01,0.2499][0.2601,0.5]	1.00	0.774	0.90	0.627
Size	F	[0.05,0.0667][0.0833,0.1]	1.00	0.879	1.00	0.815
	G	[0.05,0.0745][0.0755,0.1]	1.00	0.616	1.00	0.670
Lateral Friction	F	[0.1,0.4][0.7,1.0]	1.00	0.949	1.00	0.889
	G	[0.1,0.5406][0.5594,1.0]	1.00	0.630	1.00	0.672
Spinning Friction	F	[0.001,0.334][0.667,1.0]	1.00	0.846	1.00	0.914
	G	[0.001,0.4901][0.5109,1.0]	0.95	0.600	1.00	0.616
Gravity	F	[-1.0,-4.5][-8.0,-11.5]	1.00	0.896	1.00	0.912
	G	[-1.0,-6.1406][-6.3594,-11.5]	1.00	0.767	1.00	0.669

Table 4: Analysis of gap size effect in the estimation of causal factors

of the causal factors (Spinning Friction and Mass), the identification of the factors is not always correct, obtaining a slightly lower F1 score, and in general, we observe in all of the factors a significant drop in the clustering score values, decreasing on average very similarly for the CEM and the PPO planners by 24.4% and 25.3% respectively. This is consistent with the results of the previous experiment in which we observed that the lower

dispersion of the clusters improved the accuracy of the factor determination. It is important to note that although a lower clustering score may indicate a lower robustness of the results, in this experiment we are still able to identify correctly the causal factors in almost all cases. Therefore, although we begin to observe the limitations of this methodology, it is still satisfactory for the design of an effective exploration strategy.

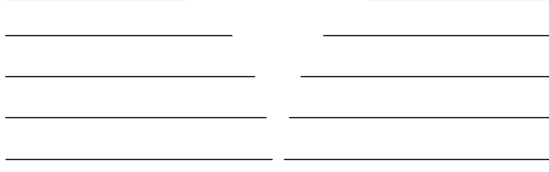


Fig. 6: Reduction of gap size

4.4 Multiple causal factors estimation

In this section, we investigate the effect of simultaneously modifying several causal factors (RQ4). So far we have considered ideal experimental scenarios, in which we can make interventions on one variable while holding all other variables constant. Here we investigate less gracious real-world scenarios where this is not possible. In this case, two causal factors vary. Our objective is to determine the value of one of them (the main causal factor), which continues to vary between two ranges of values as in the previous experiments. But now, in each of these two groups of scenarios, there is additionally a variation of the other factor (the secondary causal factor) taking place as well between two ranges of values. E.g. the two groups to be clustered and identified as different could be ‘high mass with large and small sizes’ versus ‘low mass with large and small sizes’. In Figure 7 we represent this experiment. To correctly understand this scenario we highlight that the two factors do not vary in a correlated way (in the previous example, it is not ‘high mass and large size’ versus ‘low mass and small size’). Instead, the secondary factor can be seen as a noise effect that makes it difficult to determine the first one. The variation of the second causal factor is randomised between its two possible ranges, but maintains the same order with respect to each main causal factor and also with respect to all the exploration strategies in each experiment.

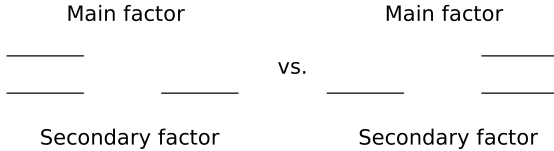


Fig. 7: Variation of two causal factors

We report the results in Tables 5, 6, 7 and 8. In each table, the first column represents the main causal factor, and the first row the secondary causal factor. We have included all combinations, except for Mass and Gravity whose effects partially counteract each other. For clarity, we present F1 and clustering scores separately. In this case, we do not observe a general trend for all the factors that allow us to quantify the overall improvement or worsening, but rather we obtain qualitative results depending on the combination of factors to be studied.

It can be seen from the F1 score results that in several of these experiments, it is not possible to correctly identify the main causal factor. We are confronted with the first group of experiments that clearly reflect the limitations of the methodology used.

We also observe that the ability to identify factors changes significantly depending on the chosen factor, something that has also not been observed so far in the previous experiments. In particular, we see for example how it is very difficult to determine the Spinning Friction value, while it is easy to identify the Size independently of the noise in the other factors. The latter makes sense, as dynamics that distinguish between different sizes can be determined mainly by the point where the force is applied, which is independent of the other factors, while in other cases, the interaction of the two factors may be correlated (at least depending on the specific type of movement triggered by each exploration strategy). We also note that Lateral and Spinning Friction are both less relevant as secondary factors.

It is therefore relevant when studying the dynamics of a particular system to take into account the specific hierarchy that can be established between the different factors, in order to be able to develop suitable experimentation strategies for their identification.

In this case, it should be noted that the clustering score should be considered in a secondary way with respect to the F1 score. That is, a high clustering score indicates clusters with little spread, but it is still of little relevance if part of the elements of each cluster are erroneous and belong to the opposite cluster. Nevertheless, this score is useful to see some of the previous cases in more detail. For example, Mass can be correctly identified as the main factor with respect to Lateral or

	Mass	Size	Lat. Frict.	Spin. Frict.	Gravity
Mass		0.85	0.95	1.00	
Size	1.00		1.00	1.00	1.00
Lat. Frict.	0.85	1.00		1.00	1.00
Spin. Frict.	0.65	0.55	0.75		0.65
Gravity		0.90	1.00	1.00	

Table 5: Analysis of multiple causal factors estimation. CEM planner. F1 score.

	Mass	Size	Lat. Frict.	Spin. Frict.	Gravity
Mass		0.766	0.722	0.883	
Size	0.804		0.848	0.827	0.788
Lat. Frict.	0.474	0.641		0.832	0.877
Spin. Frict.	0.482	0.701	0.544		0.725
Gravity		0.530	0.731	0.839	

Table 6: Analysis of multiple causal factors estimation. CEM planner. Clustering score.

Spinning Friction. But in the latter case the clustering value is much higher than in the former. This is consistent with the fact that Spinning Friction obtains the lowest values of identification as a main factor.

4.5 Causally related and confounding causal factors determination

The last experiment represents the most complex scenario of this work. We allow several factors to vary during the experiment, as in the previous case, but now the factors are causally related to each other (RQ5), while in the last experiment the factors were uncorrelated. We analyse two causal models, in the first one, a factor (Gravity) is the cause of a second factor (Lateral Friction). In the second model, one factor (Gravity) is a confounding variable with respect to two other factors (Lateral and Spinning Friction). In Figure 8 we represent the Causal DAG corresponding to these two cases, where we omit the parameters that do not vary and the observations. These types of DAG represent two key scenarios with respect to the causal study of systems with two and three variables.

To analyse these causal relationships, and given the experimental framework used, we propose to model these scenarios by means of discrete

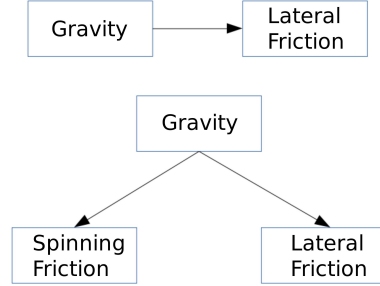


Fig. 8: Causal Directed Acyclic Graphs between the factors

Additive Noise Models (ANM) [91]. The first causal model is defined by the following equations:

$$\begin{aligned}
 \mathbf{L} &= f_L(\mathbf{G}) + \mathbf{N}_1 \text{ and } \mathbf{N}_1 \perp \mathbf{G} \\
 \mathbf{G} &\neq f_G(\mathbf{L}) + \mathbf{N}_2 \text{ and } \mathbf{N}_2 \perp \mathbf{L}
 \end{aligned} \tag{7}$$

where \mathbf{L} and \mathbf{G} represent the Gravity and Lateral Friction values, f_L and f_G are independent functions, and \mathbf{N}_n are independent noise variables for each factor.

As explicitly stated in this equation, in the case of these ANMs the above equality only holds in the causal direction.

The values of the factors corresponding to the first model are shown in Figure 9. In the left part of the figure, we see the values, where each

	Mass	Size	Lat. Frict.	Spin. Frict.	Gravity
Mass		0.95	1.00	1.00	
Size	1.00		1.00	1.00	1.00
Lat. Frict.	0.80	0.60		1.00	1.00
Spin. Frict.	0.65	0.55	1.00		0.75
Gravity		0.75	1.00	1.00	

Table 7: Analysis of multiple causal factors estimation. PPO planner. F1 score.

	Mass	Size	Lat. Frict.	Spin. Frict.	Gravity
Mass		0.608	0.615	0.807	
Size	0.797		0.800	0.811	0.743
Lat. Frict.	0.362	0.595		0.873	0.590
Spin. Frict.	0.696	0.767	0.693		0.404
Gravity		0.681	0.832	0.885	

Table 8: Analysis of multiple causal factors estimation. PPO planner. Clustering score.

square indicates that each variable does not have a point distribution but represents a range of values (the line of previous representations is now, therefore, a square). Next, we observe two possible ways of clustering these values, identifying the value of Gravity in the second figure, and of Lateral Friction in the third.

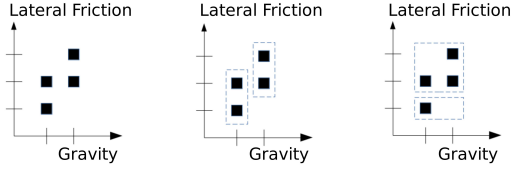


Fig. 9: Values for the first causal model, and two clustering scenarios (C1 and C2)

The second causal model is defined by the following equations:

$$\begin{aligned} \mathbf{L} &= h_L(\mathbf{G}) + \mathbf{N}_3 \text{ and } \mathbf{N}_3 \perp \mathbf{G} \\ \mathbf{S} &= h_S(\mathbf{G}) + \mathbf{N}_4 \text{ and } \mathbf{N}_4 \perp \mathbf{G} \end{aligned} \quad (8)$$

where \mathbf{S} represents the Spinning Friction, and we omit the anticausal inequalities.

This second experiment is represented in Figure 10. As we can see, this can be represented as two parallel relationships in which in both cases Gravity is the main cause of the other factors.

In this figure, we represent four possible ways of clustering the values. The first two represent the main identification of the values: Gravity in the first case and the other two factors simultaneously in the second case as they are correlated. In the last two graphs (above and below) we represent the case in which once both factors (Lateral and Spinning Friction) are identified we apply a new strategy to differentiate between one and the other, in case their dispersion is not equal.

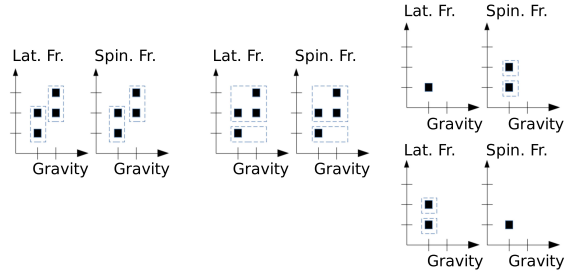


Fig. 10: Four clustering scenarios for the second causal model (C3 to C6)

The results of these experiments are presented in Table 9.

It can be seen from the results that in most cases it is possible to correctly identify the values of the causal factors. In relation to the first causal model with two variables (results C1 and C2 in

Causal factors	Range	CEM planner		PPO planner	
		F1	Clust.	F1	Clust.
C1 $G \rightarrow L$ cluster G	G:[-1.0,-4.5][-8.0,-11.5] L:[0.1,0.28][0.46,0.64][0.82,1.0]	1.00	0.848	1.00	0.903
C2 $G \rightarrow L$ cluster L	G:[-1.0,-4.5][-8.0,-11.5] L:[0.1,0.28][0.46,0.64][0.82,1.0]	1.00	0.626	1.00	0.740
C3 $G \rightarrow L, S$ cluster G	G:[-1.0,-4.5][-8.0,-11.5] L:[0.1,0.28][0.46,0.64][0.82,1.0] S:[0.001,0.2008][0.4006,0.6004] [0.8002,1.0]	1.00	0.811	1.00	0.803
C4 $G \rightarrow L, S$ cluster L, S	G:[-1.0,-4.5][-8.0,-11.5] L:[0.1,0.28][0.46,0.64][0.82,1.0] S:[0.001,0.2008][0.4006,0.6004] [0.8002,1.0]	1.00	0.621	1.00	0.698
C5 $G \rightarrow L, S$ cluster L	G:[-1.0,-4.5] L:[0.1,0.28][0.46,0.64] S:[0.001,0.2008]	1.00	0.691	1.00	0.846
C6 $G \rightarrow L, S$ cluster S	G:[-1.0,-4.5] L:[0.1,0.28] S:[0.001,0.2008][0.4006,0.6004]	0.95	0.927	1.00	0.800

Table 9: Analysis of causally related and confounding causal factors estimation

the table), we obtain a slightly lower value for the second result, compatible with the results of the previous experiment and the fact that Lateral Friction was clustered in that case. In relation to the second model, we obtain similar results both in the first factor identification (C3 and C4) and in the finer discrimination between Lateral and Spinning Friction (C5 and C6). We obtain again slightly better results using the PPO planner.

It is important to clarify that these parameter identification strategies only allow us to determine the values of the factors. To determine the causal relationships between the variables we must apply causal structure identification techniques [92–94] to the results. In this case, expert knowledge about the system could allow us to postulate an ANM, and values following a relationship similar to Figures 9 or 10 would indicate what the causal relationship is. However, we have presented a simple model in this work; in a real-world scenario, we would need a much larger collection of values to be able to determine this causal direction with certainty. Nevertheless, this simple model gives us a hopeful intuition about the scope of the Causal

Curiosity method and the possibility of tackling the problem of causal structure identification.

5 Discussion

From the results of the experiments carried out, we can begin to glimpse the potential and limitations of Causal Curiosity. In simple scenarios where it is possible to vary a single parameter independently of the others, and where the ranges of variation are well defined (sections 4.1 and 4.2), this methodology works very efficiently and robustly in estimating the parameters that define our system. Under these conditions, it is also possible to estimate the parameters with very high precision (Section 4.2). The downside is that, as in many RL experiments, the method used here may involve a high number of experiments (directly proportional to the desired level of precision). In both scenarios, it is possible to apply a simple search methodology for exploration strategies. However, in order to reduce the number of experiments, an interesting line of future work opens

up in which to design more optimal partitioning strategies than the bi-partitioning carried out here. In this experiment, we observe how we can determine the parameters in a range three orders of magnitude finer and at the same time obtain an improvement in the clustering score on average up to 14.4% for the more efficient PPO planer. One specific possibility to work on for the future line of work mentioned could be to use the clustering score as a measure of uncertainty of the parameter estimation in a Bayesian-like approach. In the methodology explored here, the range of parameter exploration is divided at each iteration in an agnostic manner, only driven by the segment offering the highest score. In this other proposed line of work, the score may serve as a weight that quantifies the certainty with which we know the parameter. This implies that sometimes it may be more relevant to explore unknown areas to reduce our lack of knowledge rather than following a greedy approach.

On the other hand, we observe the limits of the methodology in two situations. The first occurs when the ranges of the parameters to be identified are close to each other (section 4.3). Here, by reducing the gap between them to about 4% of their size, we observe a reduction in the clustering score of about 25%. This is not an unsolvable constraint, but it reveals limits to be taken into account. In this case, we can divide these wide ranges with a small gap in between into multiple well-defined sub-ranges with a wide gap (similar to what was done in Section 4.2), although this obviously increases the number of explorations to be carried out.

The second challenging situation represents a scenario in which ideal experimental conditions are not possible, and at least a second factor varies at the same time as the first (section 4.4). Here we see how this case represents a hard limit of our methodology, and even with the simple experimental scenario employed here we observe cases in which it is not possible to correctly estimate the parameters. We observe the important effect of choosing to estimate one parameter or another. This second situation is therefore totally dependent on the type of dynamic system to be investigated. It also points to a second interesting line of future research, in which to compare very different systems in order to develop robust Causal Curiosity methodologies both in relation to the

systems investigated and to the choice of causal factors. At the end of this section, we propose a number of use cases where this methodology could be applied. The diversity of domains suggested is a good example of dynamical systems governed by very different equations, which would allow a more comprehensive understanding of the effect of the combination of parameters in the methodology.

The last scenario we have analysed proposes a simple modelling of a scenario with factors with causal relationships between them (Section 4.5). This involves taking a step beyond the previous section by moving even further away from the ideal experimental conditions. We observe again a better result for the PPO planer. We should not extrapolate the conclusions of this scenario too far, again due to the dependence on system, factors, and in this case causal relationships. Nevertheless, the positive outcomes of this first experiment open a door to the possibility of being successful in other situations. Additionally, it serves as a starting point for the development of experiments where the impact of these causal relationships between factors is analysed more comprehensively. This opens a path as a third future line of research in Causal ML combining the field of Causal structure identification [92–94] with the type of Reinforcement Learning experimentation carried out in this work. The specific proposal for future work would be a methodology in which to sequentially combine parameter estimation experiments with structure identification experiments. One of the tools available in the latter domain is the study of interventions as a tool to clarify the causal relationship between parameters. In simple terms, by observing correlations between variables we can design specific experiments to differentiate correlations from true causal relationships. In this case, the value of the parameters is irrelevant, the interest is only in the relationships (the arrows in the causal diagrams). By alternating these experiments with the ones proposed in this paper, we can in a combined way increase our knowledge of both elements, the value of the parameters and their relationships, and reduce their shared interplay.

Taking these results into account, we can consider how they affect the generalisability of this methodology to other scenarios than the ones explored in this work. First, we see that it is directly generalisable to any other dynamical system in which we only need to estimate one

parameter. The methodology allows us to identify the interaction in the system that maximises the accuracy in this estimation, and to increase the precision of the value obtained, reducing the estimation range as much as required. The price to pay for increasing the accuracy is the increase in the number of interactions to be carried out. The number of interactions grows linearly with the number of times we want to subdivide the value range of a parameter. We can provide an example of a scenario where this methodology would be applicable. Let's assume the case of an autonomous vehicle driving in rough terrain. In this case, we can identify certain internal parameters of the system that are likely to be particularly affected, such as the wheel pressure, for example. The application of the methodology will produce the series of movements that will most accurately identify one pressure value or another. Independently, we can apply the methodology again to estimate another parameter, such as the orientation of the wheels relative to the vehicle axle, producing another set of movements.

The limitations observed in the last experiments show in turn the situations in which the application of the method can not be generalised. For instance, in the last example, let us consider the scenario in which we want to identify an environmental parameter external to the system such as the coefficient of friction of the ground, and also an internal parameter such as the brake wear. In this case, the two parameters are confounding variables (unlike the previous one in which the two parameters were mutually independent) so that for example when observing a prolonged displacement when braking we would not be able to identify whether the effect is produced by brake wear or a low coefficient of friction of the ground. Generalisability is therefore assured in the case of a single parameter, and in the case of multi-parameter estimation depends on the relationship between the parameters of each system. For this main limitation of the methodology, as a specific proposal for a fourth future line of work, we can consider a more advanced version of the methodology in which the reward includes a factor that precisely maximises the distinguishability between related parameters. In this case, experiments should include a minimum of two unknown parameters, but ideally more, and analyse the ability of the system to autonomously identify the

combinations that are distinguishable from those that are not. As mentioned above, this opens up a line of work that connects directly to the field of causal structure identification as a subdomain of causal analysis, which would be a very fruitful addition to our line of work.

This methodology is especially useful in cases where the user cannot directly apply a measurement tool to estimate the relevant parameter. An exemplary case is that of autonomous systems. We have mentioned the case of an autonomous driving vehicle, other cases can be UAVs and their interactions with the environment, and robotic space exploration systems such as satellites or planetary rovers. On the other hand, the methodology is also relevant in the case of existing parameter measurement tools, as it offers a transversal procedure to corroborate the value of the measurement and thus ensure its validity. This is especially important in systems that require redundancy in the estimation of their states, such as aircrafts or other aerospace applications.

6 Conclusions

In this project, we have analysed the limits and challenges presented by Causal Curiosity as a methodology to identify causal factors in a dynamic system, focusing on the case of a robotic manipulator interacting with unknown objects. We have presented different experimental scenarios to explore the effectiveness and scope of the methodology in relation to: the analysis of different types of factors, the definition of their ranges in terms of precision and granularity, and the interaction of multiple factors in an uncorrelated way or through causal relationships. We have also presented comparative results between two search methods for factor estimation strategies, with typically better results from our proposed PPO planer. The results obtained show us both the limitations and the extensive potential of this methodology, and offer us concrete proposals for designing effective exploration strategies that can be applied to more complex frameworks or real-world use cases.

Declarations

- Conflict of interest. The authors declare that there is neither funding nor conflict of interest.

- Data Availability. All data included in this study are available upon request by contact with the corresponding author.

References

- [1] Schölkopf, B.: Causality for Machine Learning, 1st edn., pp. 765–804. Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3501714.3501755>
- [2] Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. *Proceedings of the IEEE* **109**(5), 612–634 (2021) <https://doi.org/10.1109/JPROC.2021.3058954>
- [3] Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: *International Conference on Machine Learning*, pp. 4114–4124 (2019). PMLR
- [4] Suter, R., Miladinovic, D., Schölkopf, B., Bauer, S.: Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In: *International Conference on Machine Learning*, pp. 6056–6065 (2019). PMLR
- [5] Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J.: On causal and anticausal learning. In: *Proceedings of the 29th International Conference on Machine Learning. ICML’12*, pp. 459–466. Omnipress, Madison, WI, USA (2012)
- [6] Kilbertus, N., Parascandolo, G., Schölkopf, B.: Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524* (2018)
- [7] Lu, C., Schölkopf, B., Hernández-Lobato, J.M.: Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576* (2018)
- [8] Singla, S., Feizi, S.: Salient imagenet: How to discover spurious features in deep learning? In: *International Conference on Learning Representations* (2022)
- [9] Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 456–473 (2018)
- [10] Wang, Z., Xiao, X., Xu, Z., Zhu, Y., Stone, P.: Causal dynamics learning for task-independent state abstraction. In: *International Conference on Machine Learning*, pp. 23151–23180 (2022). PMLR
- [11] De Haan, P., Jayaraman, D., Levine, S.: Causal confusion in imitation learning. *Advances in neural information processing systems* **32** (2019)
- [12] Ortega, P.A., Kunesch, M., Delétang, G., Genewein, T., Grau-Moya, J., Veness, J., Buchli, J., Degraeve, J., Piot, B., Perolat, J., et al.: Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819* (2021)
- [13] Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.-S., Wen, J.-R.: Counterfactual VQA: A cause-effect look at language bias. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12700–12710 (2021)
- [14] Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., Chang, K.-W.: Generating natural language adversarial examples. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2890–2896. Association for Computational Linguistics, Brussels, Belgium (2018). <https://doi.org/10.18653/v1/D18-1316>
- [15] Chen, Y., Zhang, Y., Bian, Y., Yang, H., KAILI, M., Xie, B., Liu, T., Han, B., Cheng, J.: Invariance principle meets out-of-distribution generalization on graphs. In: *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability* (2022)

- [16] Feng, F., Huang, W., He, X., Xin, X., Wang, Q., Chua, T.-S.: Should graph convolution trust neighbors? a simple causal inference method. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21, pp. 1208–1218. Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3404835.3462971>
- [17] Pearl, J.: Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press, Cambridge (2009)
- [18] Glymour, M., Pearl, J., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
- [19] Peters, J., Janzing, D., Schölkopf, B.: Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press (2017)
- [20] Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, prediction, and search. MIT press (2000)
- [21] Sontakke, S.A., Mehrjou, A., Itti, L., Schölkopf, B.: Causal curiosity: RL agents discovering self-supervised experiments for causal representation learning. In: International Conference on Machine Learning, vol. 139, pp. 9848–9858 (2021). PMLR
- [22] Thompson, J., Kasun Prasanga, D., Murakami, T.: Identification of unknown object properties based on tactile motion sequence using 2-finger gripper robot. Precision Engineering **74**, 347–357 (2022) <https://doi.org/10.1016/j.precisioneng.2021.12.009>
- [23] MKC, D.C., Shimono, T.: Inertia compensation of motion copying system for dexterous object handling. IEEEJ Journal of Industry Applications **7**(6), 495–505 (2018) <https://doi.org/10.1541/ieejia.7.495>
- [24] Murali, A., Li, Y., Gandhi, D., Gupta, A.: Learning to grasp without seeing. In: Xiao, J., Kröger, T., Khatib, O. (eds.) Proceedings of the 2018 International Symposium on Experimental Robotics, pp. 375–386. Springer, Cham (2020)
- [25] Yuan, W., Zhu, C., Owens, A., Srinivasan, M.A., Adelson, E.H.: Shape-independent hardness estimation using deep learning and a gelsight tactile sensor. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 951–958 (2017). <https://doi.org/10.1109/ICRA.2017.7989116>
- [26] Yi, Z., Calandra, R., Veiga, F., Hoof, H., Hermans, T., Zhang, Y., Peters, J.: Active tactile object exploration with gaussian processes. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4925–4930 (2016). <https://doi.org/10.1109/IROS.2016.7759723>
- [27] Kaboli, M., Feng, D., Yao, K., Lanillos, P., Cheng, G.: A tactile-based framework for active object learning and discrimination using multimodal robotic skin. IEEE Robotics and Automation Letters **2**(4), 2143–2150 (2017) <https://doi.org/10.1109/LRA.2017.2720853>
- [28] Yu, W., Tan, J., Liu, C.K., Turk, G.: Preparing for the unknown: Learning a universal policy with online system identification. arXiv preprint arXiv:1702.02453 (2017)
- [29] Murooka, M., Nozawa, S., Kakiuchi, Y., Okada, K., Inaba, M.: Feasibility evaluation of object manipulation by a humanoid robot based on recursive estimation of the object’s physical properties. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 4082–4089 (2017). <https://doi.org/10.1109/ICRA.2017.7989469>
- [30] Mavrakis, N., Stolkin, R.: Estimation and exploitation of objects’ inertial parameters in robotic grasping and manipulation: A survey. Robotics and Autonomous Systems **124**, 103374 (2020) <https://doi.org/10.1016/j.robot.2019.103374>
- [31] Schulman, J., Wolski, F., Dhariwal, P.,

- Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
- [32] De Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. *Annals of operations research* **134**, 19–67 (2005)
- [33] Camacho, E.F., Alba, C.B.: *Model predictive control*. Springer (2013)
- [34] Kaddour, J., Lynch, A., Liu, Q., Kusner, M.J., Silva, R.: Causal machine learning: A survey and open problems. arXiv preprint arXiv:2206.15475 (2022)
- [35] Bareinboim, E.: Towards Causal reinforcement learning, ICML tutorial (2020). <https://crl.causallai.net/>
- [36] Zeng, Y., Cai, R., Sun, F., Huang, L., Hao, Z.: A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21 (2024) <https://doi.org/10.1109/TNNLS.2024.3403001>
- [37] Grimbly, S.J., Shock, J., Pretorius, A.: Causal multi-agent reinforcement learning: Review and open problems. arXiv preprint arXiv:2111.06721 (2021)
- [38] Weichwald, S., Mogensen, S.W., Lee, T.E., Baumann, D., Kroemer, O., Guyon, I., Trimpe, S., Peters, J., Pfister, N.: Learning by doing: Controlling a dynamical system using causality, control, and reinforcement learning. In: *NeurIPS 2021 Competitions and Demonstrations Track*, vol. 176, pp. 246–258 (2022). PMLR
- [39] Bareinboim, E., Forney, A., Pearl, J.: Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems* **28** (2015)
- [40] Gershman, S.J.: Reinforcement learning and causal models. *The Oxford handbook of causal reasoning* **1**, 295 (2017)
- [41] Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., Hughes, E., Battaglia, P., Botvinick, M., Kurth-Nelson, Z.: Causal reasoning from meta-reinforcement learning. arXiv preprint arXiv:1901.08162 (2019)
- [42] Amin, S., Gomrokchi, M., Satija, H., Hoof, H., Precup, D.: A survey of exploration methods in reinforcement learning. arXiv preprint arXiv:2109.00157 (2021)
- [43] Peng, S., Hu, X., Zhang, R., Tang, K., Guo, J., Yi, Q., Chen, R., Zhang, X., Du, Z., Li, L., Guo, Q., Chen, Y.: Causality-driven hierarchical structure discovery for reinforcement learning. *Advances in Neural Information Processing Systems* **35**, 20064–20076 (2022)
- [44] Rezende, D.J., Danihelka, I., Papamakarios, G., Ke, N.R., Jiang, R., Weber, T., Gregor, K., Merzic, H., Viola, F., Wang, J., et al.: Causally correct partial models for reinforcement learning. arXiv preprint arXiv:2002.02836 (2020)
- [45] Pitis, S., Creager, E., Garg, A.: Counterfactual data augmentation using locally factored dynamics. *Advances in Neural Information Processing Systems* **33**, 3976–3990 (2020)
- [46] Seitzer, M., Schölkopf, B., Martius, G.: Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems* **34**, 22905–22918 (2021)
- [47] Molina, A.M., Avelino, I.F., Morales, E.F., Sucar, L.E.: Causal based Q-learning. *Research in Computing Science* **149**, 95–104 (2020)
- [48] Pathak, D., Agrawal, P., Efros, A.A., Darrell, T.: Curiosity-driven exploration by self-supervised prediction. In: *International Conference on Machine Learning*, vol. 70, pp. 2778–2787 (2017). PMLR
- [49] Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., Efros, A.A.: Large-scale study of curiosity-driven learning. In: *International Conference on Learning Representations* (2019)

- [50] Schmidhuber, J.: Curious model-building control systems. In: Proceedings 1991 IEEE International Joint Conference on Neural Networks, pp. 1458–14632 (1991). <https://doi.org/10.1109/IJCNN.1991.170605>
- [51] Chentanez, N., Barto, A., Singh, S.: Intrinsically motivated reinforcement learning. *Advances in neural information processing systems* **17** (2004)
- [52] Lehman, J., Stanley, K.O., *et al.*: Exploiting open-endedness to solve problems through the search for novelty. In: ALIFE, pp. 329–336 (2008)
- [53] Oudeyer, P.-Y., Kaplan, F.: What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurobotics* **1** (2007) <https://doi.org/10.3389/neuro.12.006.2007>
- [54] Sun, Y., Gomez, F., Schmidhuber, J.: Planning to be surprised: Optimal bayesian exploration in dynamic environments. In: Schmidhuber, J., Thórisson, K.R., Looks, M. (eds.) *Artificial General Intelligence*, pp. 41–51. Springer, Berlin, Heidelberg (2011)
- [55] Still, S., Precup, D.: An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences* **131**, 139–148 (2012)
- [56] Baldassarre, G., Mirolli, M.: *Intrinsically motivated learning in natural and artificial systems*. Springer (2013)
- [57] Baranes, A., Oudeyer, P.-Y.: Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems* **61**(1), 49–73 (2013) <https://doi.org/10.1016/j.robot.2012.05.008>
- [58] Barto, A.G.: In: Baldassarre, G., Mirolli, M. (eds.) *Intrinsic Motivation and Reinforcement Learning*, pp. 17–47. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-32375-1_2
- [59] Stadie, B.C., Levine, S., Abbeel, P.: Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814* (2015)
- [60] Mohamed, S., Jimenez Rezende, D.: Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems* **28** (2015)
- [61] Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., Abbeel, P.: VIME: Variational information maximizing exploration. *Advances in neural information processing systems* **29** (2016)
- [62] Osband, I., Blundell, C., Pritzel, A., Van Roy, B.: Deep exploration via bootstrapped DQN. *Advances in neural information processing systems* **29** (2016)
- [63] Forestier, S., Portelas, R., Mollard, Y., Oudeyer, P.-Y.: Intrinsically motivated goal exploration processes with automatic curriculum learning. *Journal of Machine Learning Research* **23**(152), 1–41 (2022)
- [64] Tang, H., Houthooft, R., Foote, D., Stooke, A., Xi Chen, O., Duan, Y., Schulman, J., DeTurck, F., Abbeel, P.: #exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems* **30** (2017)
- [65] Colas, C., Sigaud, O., Oudeyer, P.-Y.: Gexp: Decoupling exploration and exploitation in deep reinforcement learning algorithms. In: *International Conference on Machine Learning*, vol. 80, pp. 1039–1048 (2018). PMLR
- [66] Laversanne-Finot, A., Pere, A., Oudeyer, P.-Y.: Curiosity driven exploration of learned disentangled goal spaces. In: *Conference on Robot Learning*, vol. 87, pp. 487–504 (2018). PMLR
- [67] Oudeyer, P.-Y.: Computational theories of curiosity-driven learning. *arXiv preprint arXiv:1802.10546* (2018)
- [68] Guo, Z., Thakoor, S., Pîslar, M., Avila Pires,

- B., Altché, F., Tallec, C., Saade, A., Calandriello, D., Grill, J.-B., Tang, Y., *et al.*: BYOL-Explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems* **35**, 31855–31870 (2022)
- [69] Gasse, M., Grasset, D., Gaudron, G., Oudeyer, P.-Y.: Causal reinforcement learning using observational and interventional data. *arXiv preprint arXiv:2106.14421* (2021)
- [70] Ke, N.R., Didolkar, A.R., Mittal, S., Goyal, A., Lajoie, G., Bauer, S., Rezende, D.J., Bengio, Y., Pal, C., Mozer, M.C.: Systematic evaluation of causal discovery in visual model based reinforcement learning. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021)
- [71] Nair, S., Zhu, Y., Savarese, S., Fei-Fei, L.: Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751* (2019)
- [72] Ding, W., Lin, H., Li, B., Zhao, D.: Generalizing goal-conditioned reinforcement learning with variational causal reasoning. *Advances in Neural Information Processing Systems* **35**, 26532–26548 (2022)
- [73] Thomas, V., Pondard, J., Bengio, E., Sarfati, M., Beaudoin, P., Meurs, M.-J., Pineau, J., Precup, D., Bengio, Y.: Independently controllable factors. *arXiv preprint arXiv:1708.01289* (2017)
- [74] Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in beta-VAE. *arXiv preprint arXiv:1804.03599* (2018)
- [75] Kim, H., Mnih, A.: Disentangling by factorising. In: *International Conference on Machine Learning*, vol. 80, pp. 2649–2658 (2018). PMLR
- [76] Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems* **31** (2018)
- [77] Volodin, S., Wichers, N., Nixon, J.: Resolving spurious correlations in causal models of environments via interventions. *arXiv preprint arXiv:2002.05217* (2020)
- [78] Coumans, E., *et al.*: Bullet real-time physics simulation. URL <http://bulletphysics.org> (2013)
- [79] Wuthrich, M., Widmaier, F., Grimminger, F., Joshi, S., Agrawal, V., Hammoud, B., Khadiv, M., Bogdanovic, M., Berenz, V., Viereck, J., *et al.*: Trifinger: An open-source robot for learning dexterity. In: *Conference on Robot Learning*, vol. 155, pp. 1871–1882 (2021). PMLR
- [80] Sutton, R.S., Barto, A.G.: *Reinforcement learning: An introduction*. MIT press (2018)
- [81] Rissanen, J.: Modeling by shortest data description. *Automatica* **14**(5), 465–471 (1978) [https://doi.org/10.1016/0005-1098\(78\)90005-5](https://doi.org/10.1016/0005-1098(78)90005-5)
- [82] Grünwald, P.D., *et al.*: *The minimum description length principle*. MIT Press Books 1 (2007)
- [83] 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) [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- [84] Cuturi, M., Blondel, M.: Soft-DTW: a differentiable loss function for time-series. In: *International Conference on Machine Learning*, vol. 70, pp. 894–903 (2017). PMLR
- [85] Bagirov, A.M., Aliguliyev, R.M., Sultanova, N.: Finding compact and well-separated clusters: Clustering using silhouette coefficients. *Pattern Recognition* **135**, 109144 (2023) <https://doi.org/10.1016/j.patcog.2022.109144>
- [86] Punhani, A., Faujdar, N., Mishra, K.K., Subramanian, M.: Binning-based

- silhouette approach to find the optimal cluster using K-means. *IEEE Access* **10**, 115025–115032 (2022)
<https://doi.org/10.1109/ACCESS.2022.3215568>
- [87] Shutaywi, M., Kachouie, N.N.: Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy* **23**(6) (2021)
<https://doi.org/10.3390/e23060759>
 - [88] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D.: Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence* **32**(1) (2018)
<https://doi.org/10.1609/aaai.v32i1.11694>
 - [89] Eimer, T., Lindauer, M., Raileanu, R.: Hyperparameters in reinforcement learning and how to tune them. In: *International Conference on Machine Learning*, vol. 202, pp. 9104–9149 (2023). PMLR
 - [90] Kiran, M., Ozyildirim, M.: Hyperparameter tuning for deep reinforcement learning applications. *arXiv preprint arXiv:2201.11182* (2022)
 - [91] Peters, J., Janzing, D., Scholkopf, B.: Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(12), 2436–2450 (2011)
<https://doi.org/10.1109/TPAMI.2011.71>
 - [92] Vowels, M.J., Camgoz, N.C., Bowden, R.: D’ya like DAGs? a survey on structure learning and causal discovery. *ACM Comput. Surv.* **55**(4) (2022)
<https://doi.org/10.1145/3527154>
 - [93] Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. *Frontiers in Genetics* **10** (2019)
<https://doi.org/10.3389/fgene.2019.00524>
 - [94] Hasan, U., Hossain, E., Gani, M.O.: A survey on causal discovery methods for temporal and non-temporal data. *arXiv preprint arXiv:2303.15027* (2023)