Differentiable Causal Search

Kaveh Aryan KAVEH.ARYAN@KCL.AC.UK
Hana Chockler HANA.CHOCKLER@KCL.AC.UK
Mohammad Reza Mousavi MOHAMMAD.MOUSAVI@KCL.AC.UK
King's College London, London, UK

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Actual causality–identifying the causes of particular events–is formalised by the Halpern–Pearl (HP) definitions via counterfactual reasoning over structural causal models. Computing HP causes requires solving a combinatorial optimisation problem that is, depending on the variant, D_1^P-complete or worse. We propose a differentiable approximation of HP causality that leverages the robustness semantics of logical specifications and additive intervention relaxations. Specifically, we replace discrete satisfiability constraints with continuous robustness scores, and model interventions as soft variable shifts rather than hard graph surgeries. This, along with a sparsity relaxation, allows for using continuous optimisation techniques such as gradient descent. Experiments on synthetic graphs show that our method, on average, approximates the true causes with a ±5 % error margin, while achieving at least a 60× speedup. The framework also supports fine-grained control over additional causal properties such as the desired counterfactual robustness.

Keywords: actual causality, gradient descent, optimisation

1. Introduction

Actual causality aims to pinpoint the factors that led to a given outcome with applications in domains as diverse as explainable AI (Carloni et al., 2023) and legal reasoning (Moore, 2019). Unlike general (or type-level) causality, which concerns statistical or generic causal dependencies, actual causality seeks to identify the particular factors that were responsible for a given event in a given context–often within complex, high-stakes systems. A widely adopted formalisation of actual causality is the Halpern–Pearl (HP) framework (Halpern and Pearl, 2005; Halpern, 2015), which defines causality within structural causal models (SCMs) (Pearl, 2009) using three ingredients: (i) counterfactual dependence of the outcome on the candidate cause, (ii) a suitable *contingency*–a setting of other variables that makes the counterfactual dependence manifest, and (iii) minimality to exclude superfluous components from the cause. This framework has proven highly useful and applicable across a wide range of problems in computer science (Meliou et al., 2010; Sarda Gou et al., 2023; Beer et al., 2012; Dubslaff et al., 2024; Chockler et al., 2021; Pouget et al., 2021; Chockler and Halpern, 2004; Beckers et al., 2022, 2023).

Despite its principled foundation, the HP framework faces two interrelated challenges that hinder its scalability and broader applicability. First, deciding HP causality is D_1^P- (Halpern, 2015) or D_2^P-complete (Aleksandrowicz et al., 2017), depending on the specific definition, and no polynomial-time algorithm is known for *actual causal search* (i.e., identifying the HP cause of a given event). Second, perhaps as a result of the first problem, HP causality is customarily applied to Boolean SCMs. This includes, for instance, all motivating examples in the original papers (Halpern and Pearl, 2001, 2005; Halpern, 2015) and all 37 examples collected in a benchmark in

a recent work (Ibrahim and Pretschner, 2020). Still, many practical systems, such as hybrid and cyber-physical systems, operate over continuous state spaces. In such settings, specifications (e.g., safety constraints) are often expressed as Boolean formulas over continuous variables. Nevertheless, the definition of HP cause is still applicable and actual causal search is still useful, for example, in root cause analysis and causal explanations.

Contribution We propose a differentiable approximation of HP actual causality tailored to continuous SCMs. Our approach incorporates three key ideas: (i) *robustness semantics*, which replace binary formula satisfaction with a continuous measure of how strongly a counterfactual intervention satisfies or violates a constraint; (ii) *soft interventions*, which generalise Pearl's surgical interventions by an additive intervention; and (iii) *sparsity relaxation*, which encourages minimal causal explanations via ℓ_1 penalties in place of combinatorial subset selection. This novel formulation transforms causal search into a smooth optimisation problem, enabling the use of gradient-based solvers. Our method preserves the core properties of HP causality while scaling to high-dimensional, continuous systems. We demonstrate its effectiveness on two synthetic benchmarks. Our code and experimental results are available on GitHub.

The remainder of the paper is structured as follows: Section 2 reviews related work; Section 3 presents the background concepts and our approximation of causal search as a continuous optimisation problem; Section 4 reports our empirical evaluation; and Section 5 concludes the paper. Further empirical evaluation is provided in the supplementary material, and all Python code used in our experiments is included therein. The associated GitHub repository is currently private to preserve anonymity and will be made public upon publication.

2. Related work

Halpern–Pearl actual causality Among the various definitions of actual causality (Mackie, 1965; Wright, 2011), the formalisation by Halpern and Pearl (Halpern and Pearl, 2005; Halpern, 2015) has gained adoption in computer science, in explainable AI (Chockler et al., 2021), software testing (Beer et al., 2012), responsibility allocation (Chockler and Halpern, 2004), and more (Beckers et al., 2023; Meliou et al., 2010). This definition is based on counterfactual dependence (Norton and Norton, 2011) of the effect (a well-formed Boolean formula) on the causal variables as expressed with the language of SCMs (Pearl, 2009). Most of the existing work focuses on systems with discrete variables (Ibrahim and Pretschner, 2020; Sarda Gou et al., 2023) with only a few addressing continuous variables (Peters and Halpern, 2021). It has been shown that deciding actual causality is at least D_1^P-complete (Aleksandrowicz et al., 2017; Halpern, 2015) leading to the recent suggestion to frame the problem of *actual causal search* in Boolean systems as a combinatorial optimisation problem and use approximation approaches such as MAX-SAT (Ibrahim and Pretschner, 2020).

Robustness semantics of logical specifications The *robustness semantics* of logical specifications extends traditional Boolean semantics by assigning a quantitative measure, known as *robustness degree*, that captures how strongly a system trajectory satisfies or violates a given specification (Fainekos and Pappas, 2009). This concept has been applied across various domains, including reinforcement learning (Aksaray et al., 2016), hybrid systems (Donzé and Maler, 2010), and motion planning (Fainekos, 2008), enabling more resilient verification, control, and synthesis under uncertainty. We use robustness semantics to relax the Boolean effect formula (events) by a continuous quantification which allows continuous optimisation approaches to causal search.

Soft interventions The concept of actual causality is formalised by the notion of interventions, or so-called graph surgeries (Pearl, 2009), where a structural equation is replaced by a new, counterfactual, constant term (Halpern, 2016). The binary notion of hard intervention is a second obstacle to the applicability of continuous optimisation methods and, thus, requires relaxation. *Soft intervention* generalises the notion to cases where the structural equation is replaced with a new function (Eberhardt and Scheines, 2007) and has been applied in the context of causal discovery (i.e., graph learning) (Jaber et al., 2020), intervention learning (Zhang et al., 2021; Besserve and Schölkopf, 2022), and abstraction (Massidda et al., 2023). In this work, our soft interventions take a particular form of addition of a constant vector to the nodes' current values. This relaxation facilitates differentiable optimisation while preserving the causal semantics necessary for actual causal analysis.

3. Method

In this section, we first present the theoretical background to our work, i.e., the Halpern–Pearl definition of actual causality and the robustness semantics of logical specifications. We then proceed to present our differentiable approach.

Note on notation. We use square brackets for three related purposes: to denote interventions (defined later), e.g., $\mathcal{M}[\mathcal{X} = \boldsymbol{x}]$ and $\mathcal{M}[\mathcal{X} = \boldsymbol{x}, \mathcal{Y} = \boldsymbol{y}]$; to denote the state induced by a context, e.g., $\boldsymbol{v} = \mathcal{M}[\mathcal{U} = \boldsymbol{u}]$; and to denote projection to a specified subset of variables, e.g., $\boldsymbol{v}[\mathcal{X}]$. In each case, the intended meaning is clear from the form of the expression and the surrounding context.

3.1. Definitions

The notion of *structural causal models* (Pearl, 2009) is standard in the formalisations of actual causality (Halpern, 2016; Beckers and Vennekens, 2018):

Definition 1 *A* structural causal model (SCM) *is a tuple* $\mathcal{M} = (\mathcal{U}, \mathcal{V}, R, \mathcal{E})$ *in which* $\mathcal{U}$ *and* $\mathcal{V}$ *are two disjoint sets of variables called* exogenous *and* endogenous *variables,* $R(X)$, $X \in \mathcal{U} \cup \mathcal{V}$, *is a function that determines the* ranges *of variables, and* $\mathcal{E}$ *is the set of* structural equations $\{X = f_X(Y_1, Y_2, Y_3, ..., Y_n)\}_{X \in \mathcal{V}}$ *where* $Y_i \in (\mathcal{U} \cup \mathcal{V}) - \{X\}$ *for all* $X \in \mathcal{V}$. *In this paper we assume all variables are real-valued, that is,* $R(X) = \mathbb{R}$ *for all* $X \in \mathcal{U} \cup \mathcal{V}$. *Therefore we omit* R *from the definition of SCMs for simplicity.*

The motivation for differentiating between exogenous and endogenous variables is to distinguish between variables that are the inputs of the system and those that are influenced by the inputs, respectively. Structural *equations* (or *assignments*, more precisely) specify how each endogenous variable is influenced by the exogenous and other endogenous variables.

In line with our focus on particular events and actual causality, we are interested in the particular values that variables of the system take. A particular setting for the exogenous variables is called a *context*, and a particular setting for the endogenous variables is called a *state*.

An SCM can be visualised through a graph where each variable in $\mathcal{U} \cup \mathcal{V}$ is represented by a node, and a directed edge from X to Y exists iff the value of Y depends, in at least one setting of variables, on the value of X. An SCM is acyclic if its corresponding graph is acyclic. If $\mathcal{M}$ is an acyclic SCM, a given context $\boldsymbol{u}$ uniquely determines the state of $\mathcal{M}$, which is denoted by

$\boldsymbol{v} = M[\mathcal{U} = \boldsymbol{u}]$. The value of a certain variable $X \in \mathcal{V}$ in $\boldsymbol{v}$ is denoted by $\boldsymbol{v}[X]$. In this paper, we align with the literature and focus on acyclic models (Zheng et al., 2018; Beckers et al., 2022).

An important notion for defining and quantifying causal effects is that of an *intervention*. Interventions refer to deliberately changing the value or behavior of variables to observe their subsequent effects on other variables. The classical HP formulation uses *hard interventions* (also called *surgical interventions*), where variables are set to fixed values. However, for continuous optimisation, we will also consider *soft interventions* that modify variables additively.

Definition 2 *If $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ is an SCM, $X \in \mathcal{V}$, and $x \in \mathbb{R}$, then the* hard-intervention model *$\mathcal{M}[X = x]$ is the SCM $(\mathcal{U}, \mathcal{V}, \mathcal{E}')$ in which $\mathcal{E} = \mathcal{E}'$ except that the equation $X = f_X(Y_1, \ldots, Y_n)$, $Y_i \in (\mathcal{U} \cup \mathcal{V}) - \{X\}$, is replaced with $X = x$. The hard-intervention model $\mathcal{M}[\mathcal{X} = \boldsymbol{x}]$ where $\mathcal{X} \subseteq \mathcal{V}$ and $\boldsymbol{x} \in \mathbb{R}^{|\mathcal{X}|}$ is defined similarly by replacing the equations for all variables in $\mathcal{X}$ with their corresponding values in $\boldsymbol{x}$.*

Definition 3 *If $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ is an SCM and $\boldsymbol{\delta} \in \mathbb{R}^{|\mathcal{V}|}$, then the* additive-intervention model *$\mathcal{M}_{\boldsymbol{\delta}}$ is the SCM $(\mathcal{U}, \mathcal{V}, \mathcal{E}')$ where each structural equation $X = f_X(Y_1, \ldots, Y_n)$ in $\mathcal{E}$ is replaced with $X = f_X(Y_1, \ldots, Y_n) + \delta_X$ in $\mathcal{E}'$, where δ_X is the component of $\boldsymbol{\delta}$ corresponding to variable X.*

The corresponding graph for a hard intervention $\mathcal{M}[X = x]$ is the graph for $\mathcal{M}$ after a so-called *graph surgery* where incoming edges for node X are removed. For additive interventions, the graph structure remains unchanged, but each variable receives an additional exogenous input. As we will see, the HP causal search is effectively a search over the space of interventions with certain properties.

Next, we present the Halpern–Pearl (HP) definition of actual causality (Halpern, 2015, 2016). We employ the latest version of the definition, i.e., the *modified* definition (Halpern, 2015).

Definition 4 *If $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ is an SCM, $\Sigma_M = (\mathcal{V}, \mathscr{C}, \mathscr{F}, \mathscr{P})$ is called a* language signature *over $\mathcal{M}$, where $\mathcal{V}$ is the set of endogenous variables in $\mathcal{M}$, $\mathscr{C} = \mathbb{R}$ is the set of constant symbols, $\mathscr{F}$ is a set of function symbols including those appearing in the structural equations, and $\mathscr{P}$ is a set of predicate symbols including equality. An* event *is a Boolean well-formed formula over the signature Σ_M, logical connectives, and quantifiers. We denote an event by φ or $\varphi(\mathcal{V})$ to emphasise that free variables of the event are in $\mathcal{V}$.*

Definition 5 *If $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ is an SCM and $\varphi(\mathcal{V})$ is an event, then $\mathcal{X} = \boldsymbol{x}$ is called an HP cause of φ, iff:*

- *HP1. $\mathcal{M}[\mathcal{U} = \boldsymbol{u}][\mathcal{X}] = \boldsymbol{x}$ and $\varphi(\mathcal{M}[\mathcal{U} = \boldsymbol{u}])$,*
- *HP2. $\exists \mathcal{W} \subseteq \mathcal{V}$, $\exists \boldsymbol{x}'$ such that $\neg\varphi(\mathcal{M}[\mathcal{U} = \boldsymbol{u}, \mathcal{X} = \boldsymbol{x}', \mathcal{W} = \boldsymbol{w}^*])$, where $\boldsymbol{w}^* = \mathcal{M}[\mathcal{U} = \boldsymbol{u}][\mathcal{W}]$,*
- *HP3. $\mathcal{X}$ is minimal, in the sense that none of its subsets satisfies HP1 and HP2.*

The tuple $(\mathcal{W}, \boldsymbol{w}^*, \boldsymbol{x}')$ is called a witness *to the fact that $\mathcal{X} = \boldsymbol{x}$ is a cause of φ.*

This definition formalises basic intuitions about the notion of an actual cause: HP1 ensures that the cause and effect have actually happened. HP2 ensures that the effect is counterfactually dependent on causes (Norton and Norton, 2011), with a caveat that this dependence might not be

visible unless under a contingency where a set $\mathcal{W}$ of system variables are held at their actual values. HP3 ensures that causal sets are minimal.

The original HP definition is based on hard interventions (Definition 2). However, for practical reasons, we prefer additive interventions (Definition 3) in our differentiable approximation. The following equivalence result justifies this choice.

Theorem 6 *If $\mathcal{M}$ is an SCM, φ an event, and $\mathcal{X} = \boldsymbol{x}$ is an HP cause of φ in context $\boldsymbol{u}$, witnessed by $(\mathcal{W}, \boldsymbol{w}^*, \boldsymbol{x}')$ as in Definition 5, then there exists a vector $\boldsymbol{\delta} \in \mathbb{R}^{|\mathcal{X} \cup \mathcal{W}|}$, such that the following holds:*

1. *If $\boldsymbol{v}^* := \mathcal{M}[\mathcal{U} = \boldsymbol{u}]$ and $\boldsymbol{w}^* := \boldsymbol{v}^*[\mathcal{W}]$, then the solution $\boldsymbol{v}'$ to the SCM $\mathcal{M}_{\boldsymbol{\delta}}$ in context $\boldsymbol{u}$ satisfies:*

$$\begin{aligned} &\boldsymbol{v}'[\mathcal{X}] = \boldsymbol{x}', \\ &\boldsymbol{v}'[\mathcal{W}] = \boldsymbol{w}^*, \\ &\neg\varphi(\boldsymbol{v}'). \end{aligned}$$

2. *Conversely, for any such $\boldsymbol{\delta}$ with these properties, there exists a hard intervention on $\mathcal{X} \cup \mathcal{W}$ that achieves the same counterfactual outcome.*

Proof Given the HP witness $(\mathcal{W}, \boldsymbol{w}^*, \boldsymbol{x}')$ and $\boldsymbol{x}'$, set $S = \mathcal{X} \cup \mathcal{W}$ and let $\boldsymbol{s}$ assign $\boldsymbol{x}'$ to $\mathcal{X}$ and $\boldsymbol{w}^*$ to $\mathcal{W}$. The counterfactual world is $\mathcal{M}[\mathcal{U} = \boldsymbol{u}, S = \boldsymbol{s}]$. Because $\mathcal{M}$ is acyclic and real-valued, for any $Z \in S$ there exists $\delta_Z := s_Z - f_Z(\boldsymbol{v}^{do}[\mathrm{Pa}(Z)])$, where $\boldsymbol{v}^{do} = \mathcal{M}[\mathcal{U} = \boldsymbol{u}, S = \boldsymbol{s}]$. The additive-intervention SCM $\mathcal{M}_{\boldsymbol{\delta}}$ (with $\boldsymbol{\delta}$ supported only on S) thus yields the same solution as the hard intervention, so all three properties hold. The converse follows similarly by simulating any additive intervention with a hard intervention. ■

It should be noted that a simple minimal intervention in the HP formulation may require a complicated $\mathcal{W}$, hence, a complicated $\boldsymbol{\delta}$ in the additive-intervention formulation. Therefore, our approach approximates the HP definition by finding a minimal $\boldsymbol{\delta}$. In a generalised definition of HP causality, where the choice of witness is not cost-free, our approach can be seen as a principled way to minimise the cause and the witness.

To bridge the gap between binary events and continuous measures of causality, we employ the robustness semantics of logical specifications (Fainekos, 2008).

Definition 7 *If $\varphi(\mathcal{V})$ is an event, the* extension set *of φ, denoted by $ext(\varphi)$, is the set of points in the $\mathcal{V}$-space that satisfy φ.*

Given an event φ, we quantify the *robustness* of the system, by measuring the distance of the state of the system to the boundary of the extension. Assuming the reader is familiar with the concept of a metric space (Willard, 2012), we recall the definition of the depth of a point in a set:

Definition 8 *If $(\mathcal{G}, d)$ is a metric space, $\boldsymbol{x} \in \mathcal{G}$, and $\mathcal{H} \subseteq \mathcal{G}$, then*

- *The* distance *from $\boldsymbol{x}$ to $\mathcal{H}$ is defined as $dist_d(\boldsymbol{x}, \mathcal{H}) = \inf\{d(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{y} \in \bar{\mathcal{H}}\}$, where $\bar{\mathcal{H}}$ is the closure of $\mathcal{H}$, that is, the intersection of all closed sets containing $\mathcal{H}$; and*

- *The* depth *of* $\boldsymbol{x}$ *in* $\mathcal{H}$ *is defined as* $depth_d(\boldsymbol{x}, \mathcal{H}) = dist_d(\boldsymbol{x}, \mathcal{G} - \mathcal{H})$.
- *The* signed distance *from* $\boldsymbol{x}$ *to* S *is defined as*

$$Dist_d(\boldsymbol{x}, S) = -dist_d(\boldsymbol{x}, S) \text{ if } \boldsymbol{x} \notin S, depth_d(\boldsymbol{x}, S) \text{ if } \boldsymbol{x} \in S$$

Now, we formally define the robustness of a point with respect to a logical specification:

Definition 9 *If* $\varphi(\mathcal{V})$ *is an event and* $\boldsymbol{x} \in \mathcal{V}$ *is a point in the* $\mathcal{V}$*-space, then the robustness of* $\boldsymbol{x}$ *wrt* φ *is defined as:*

$$\xi(\boldsymbol{x}, \varphi) = Dist_d(\boldsymbol{x}, ext(\varphi))$$

3.2. Differentiable Actual Causal Search

We next reformulate the search for an actual cause in continuous SCMs as a differentiable optimisation problem, starting from the classical Halpern–Pearl definition and progressively relaxing its discrete and non-differentiable components to obtain a scalable, gradient-based approximation.

Consider a continuous SCM $\mathcal{M}$, an event $\varphi(\mathcal{V})$, and a context $\boldsymbol{u}$. We assume φ holds true under $\boldsymbol{u}$, and our goal is to identify a candidate witness $(\mathcal{W}, \boldsymbol{w}^*, \boldsymbol{x}')$ that approximates a Halpern–Pearl (HP) cause of φ.

Denote the state of the system by $\boldsymbol{v} = \mathcal{M}[\mathcal{U} = \boldsymbol{u}]$. The original HP causal search can be framed as a combinatorial optimisation problem, as inspired by (Ibrahim and Pretschner, 2020):

$$(\mathcal{Y}_{hp}, \boldsymbol{y}'_{hp}) = \arg\min_{\mathcal{Y}, \boldsymbol{y}'} |\{\boldsymbol{y}'[Y] \neq \boldsymbol{v}[Y] : Y \in \mathcal{Y}\}| \quad \text{s.t.} \quad \neg\varphi(\mathcal{M}[\mathcal{U} = \boldsymbol{u}, \mathcal{Y} = \boldsymbol{y}']) \tag{1}$$
$$\mathcal{W}_{hp} = \left\{Y \in \mathcal{Y}_{hp} : \boldsymbol{y}'_{hp}[Y] = \boldsymbol{v}[Y]\right\}, \quad \mathcal{X}_{hp} = \mathcal{Y}_{hp} \setminus \mathcal{W}_{hp}, \quad \boldsymbol{x}' = \boldsymbol{y}'_{hp}[\mathcal{X}_{hp}].$$

To enable differentiability and scalability for continuous SCMs, we introduce three relaxation steps:

Step 1: Robustness semantics First, we replace the binary constraint φ with its robustness (Fainekos and Pappas, 2009) to obtain a continuous quantitative measure $\xi(\mathcal{M}[\mathcal{U} = \boldsymbol{u}, \mathcal{Y} = \boldsymbol{y}'], \varphi)$ of how strongly the modified SCM satisfies (positive robustness) or violates (negative robustness) the effect φ. Consequently, the original combinatorial constraint $\neg\varphi$ transforms into the following continuous constraint:

$$\xi(\mathcal{M}[\mathcal{U} = \boldsymbol{u}, \mathcal{Y} = \boldsymbol{y}'], \varphi) < 0. \tag{2}$$

Step 2: Additive interventions The constraint in Equation 2 remains impractical for gradient-based optimisation, as it depends on the non-differentiable operation of hard intervention, i.e., replacing the value of a subset of variables with fixed constants. Justified by Theorem 6, we address this by introducing a differentiable, additive intervention relaxation (Eberhardt and Scheines, 2007): for each variable $Y \in \mathcal{V}$, we augment the structural equation with a learnable additive parameter δ_Y, resulting in

$$\tilde{f}_Y(\cdot) = f_Y(\cdot) + \delta_Y, \quad \delta_Y \in \mathbb{R}. \tag{3}$$

This way, the intervention vector $\boldsymbol{\delta}$ is optimised directly, allowing smooth, continuous adjustments to the system. The relaxed causal search problem thus becomes:

$$\min_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_0 \quad \text{s.t.} \quad \xi(\mathcal{M}_{\boldsymbol{\delta}}[\mathcal{U} = \boldsymbol{u}], \varphi) < 0, \tag{4}$$

where $\mathcal{M}_{\boldsymbol{\delta}}$ denotes the SCM with additive intervention $\boldsymbol{\delta}$. This formulation ensures minimality (HP3 in Definition 5) by seeking the sparsest intervention that violates (HP2 in Definition 5) the effect.

We further transform this into an unconstrained penalised problem by introducing a hinge penalty, which penalises non-violation of the effect:

$$\min_{\boldsymbol{\delta}} \Big[\|\boldsymbol{\delta}\|_0 + \lambda \left[\xi(\mathcal{M}_{\boldsymbol{\delta}}[\mathcal{U} = \boldsymbol{u}], \varphi)\right]_+ \Big], \qquad [t]_+ := \max(t, 0), \tag{5}$$

where $\lambda > 0$ controls the trade-off between sparsity and effect violation. This formulation ensures that solutions will only incur a penalty when the effect has not been counterfactually violated ($\xi \geq 0$).

Step 3: Sparsity relaxation Finally, we relax the non-differentiable ℓ_0-norm in $\|\boldsymbol{\delta}\|_0$ by substituting it with the differentiable sparsity-inducing ℓ_1-norm. The fully differentiable causal search objective is thus:

$$J_\lambda(\boldsymbol{\delta}) = \|\boldsymbol{\delta}\|_1 + \lambda \left[\xi(\mathcal{M}_{\boldsymbol{\delta}}[\mathcal{U} = \boldsymbol{u}], \varphi)\right]_+, \tag{6}$$

where $\lambda > 0$ is a hyperparameter governing the sparsity–robustness trade-off.

This formulation enables efficient solution via standard gradient-based optimisation, and crucially, it extends HP-style actual causality discovery to continuous and high-dimensional systems.

4. Empirical Results

We evaluate our differentiable HP causal search method on two synthetic benchmarks, focusing on scalability and the trade-off between approximation quality and sparsity. We compare against exact HP causal search where feasible.

4.1. Research Questions

We evaluate our differentiable HP causal search method by addressing the following research questions:

- **RQ1 (Scalability):** How does the runtime of our differentiable approach compare to exact HP causal search methods as the number of variables increases?
- **RQ2 (Approximation–Parameter Tradeoff):** How well does our differentiable approach approximate the size and robustness of true HP causes, and how does the sparsity parameter λ modulate the trade-off between minimality and robustness violation?

4.2. Evaluation Metrics

For our research questions, we employ the following metrics:

- **RQ1:** *Runtime Reduction Ratio (RRR)*, defined as the ratio of the runtime of the exact HP method to that of our differentiable approach. For large numbers of variables, when HP runtime exceeds a threshold, we report the actual runtime.
- **RQ2:** We use the following ratios:

- *Intervention Sparsity Ratio (ISR):* the ratio of the L1 norm of intervention parameters found by the differentiable method to that found by the exact HP method, $\text{ISR} = \frac{\|\boldsymbol{\delta}_{\text{diff}}\|_1}{\|\boldsymbol{\delta}_{\text{HP}}\|_1}$.
- *Robustness*: the robustness value achieved by the differentiable method, $\xi(\mathcal{M}_{\boldsymbol{\delta}_{\text{diff}}}[\mathcal{U} = \boldsymbol{u}], \varphi)$. We report absolute robustness rather than a ratio because HP achieves perfect robustness ($\xi = 0$), making a ratio undefined.

Metric Interpretation. The *Runtime Reduction Ratio (RRR)* measures the speedup achieved by the differentiable approach over the exact HP method; higher values indicate greater runtime improvement and are therefore better. The *Intervention Sparsity Ratio (ISR)* compares the sparsity of interventions: values close to 1 indicate that the differentiable method finds interventions of comparable sparsity to the exact HP method. Values greater than 1 indicate less sparse (less minimal) interventions, while values below 1 can occur when λ is sufficiently small such that the optimizer favors very sparse interventions at the expense of non-significant robustness violation–thus, lower ISR indicates stronger sparsity but must be interpreted jointly with robustness. The *Robustness* compares robustness violations with the perfect robustness of the HP methods (that is $\xi = 0$). As a general rule, by increasing λ, we expect to see lower robustness (closer to 0) at the expense of higher ISR (less sparse interventions).

4.3. Baselines

We use exact HP causal search as our primary baseline, implemented using Breadth First Search (BFS) (Cormen et al., 2022) over all possible subsets of variables. Since there are no established baselines, we compare against this exact method where computationally feasible. We search by gradually increasing the size of candidate causes, terminating when a cause is found, which guarantees minimality (HP3 in Definition 5). At each candidate cause, we check HP2 by searching over all possible contingency sets, and use the Z3 SMT solver (de Moura and Bjørner, 2008) to find a counterfactual condition.

4.4. Methodology

We generate synthetic SCMs using two well-established graph models: Erdős–Rényi random graphs for modeling non-heterogeneous networks (Erdős and Rényi, 2022) and Barabási-Albert graphs for scale-free networks (Albert and Barabási, 2002). To ensure acyclicity, we convert these undirected graphs to directed acyclic graphs (DAGs) using topological sorting, resulting in strictly lower-triangular binary adjacency matrices.

For each DAG, we construct a random cubic polynomial SCM by assigning real-valued coefficients to the adjacency matrix, drawing independently from $\mathcal{N}(0, \mu)$, $\mu = 1, 0.05$. We choose cubic polynomials to introduce non-linearity while maintaining numerical stability. Exogenous noise terms, $\boldsymbol{u}$, are sampled from $\mathcal{U}(2, 4)$, and the factual state, $\boldsymbol{v}_a$, is computed accordingly.

We chose the last node in the adjacency matrix as the output variable, O, and define the effect formula and robustness as:

$$\varphi : O > o_{thr}, \qquad \xi(\boldsymbol{v}, \varphi) = \boldsymbol{v}[O] - o_{thr},$$

where o_{thr} is chosen as a fraction of the actual robustness, $\xi_a = \xi(\boldsymbol{v}_a, \varphi)$.

To answer RQ1, we empirically determined the maximum number of variables for which the exact HP method completes within a timeout of 60 seconds, finding it to be $n = 16$ for both graphs.

We evaluated on SCMs with $n = 10, \ldots, 16$ nodes to measure RRR. Beyond this, we report actual runtimes for $n = 50, 100, 150,$ and 200 nodes.

For RQ2, we report SCMs with $n = 10$ and 16 nodes, varying the sparsity parameter λ over $[0.01 - 2]$ smoothly to observe its effect on ISR and robustness.

4.5. Results

Figures 1 and 2 show the RRR plots for Erdős–Rényi and Barabási-Albert graphs, respectively. Our differentiable approach consistently outperforms the exact HP method. The exact runtime when the number of variables exceeds 16 is shown in Figure 3, which shows that the runtime of our method is on the order of hundredths of a second even for 200 variables, while the exact method becomes intractable.

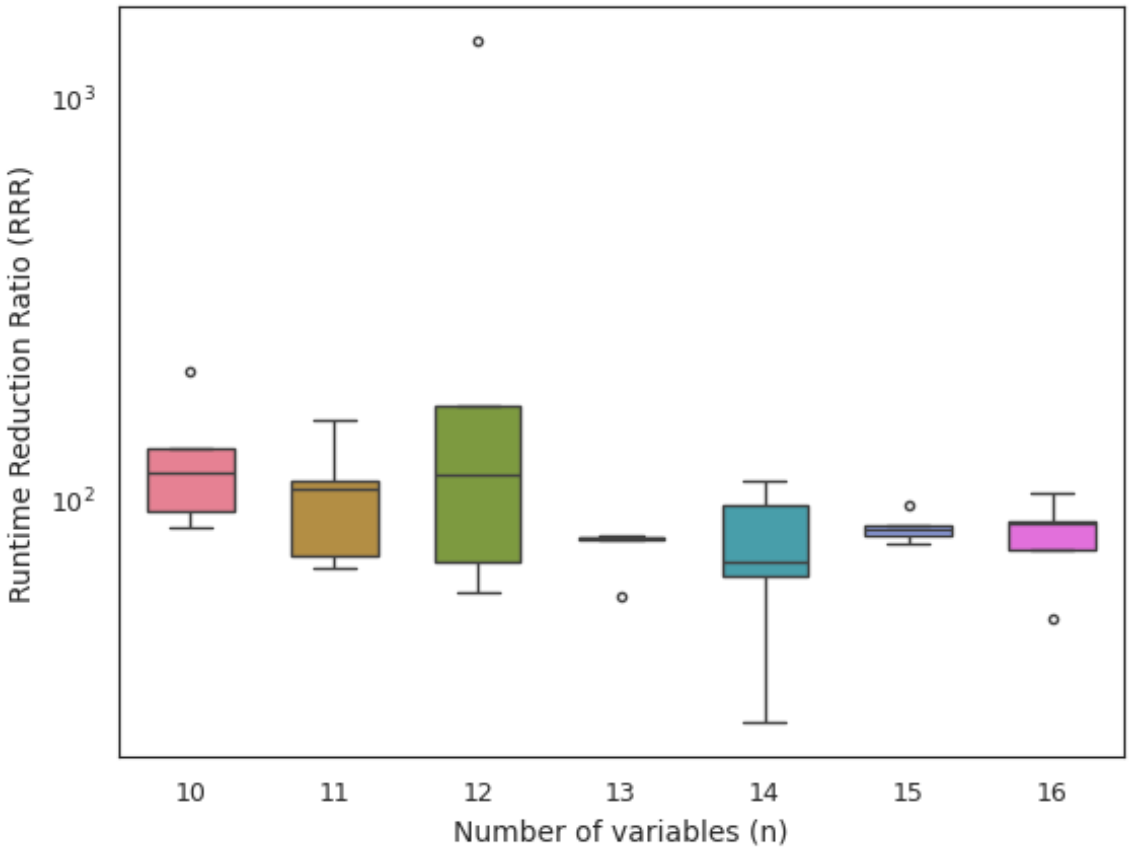

Figure 1: Runtime Reduction Ratio (RRR) for Erdős–Rényi graphs.

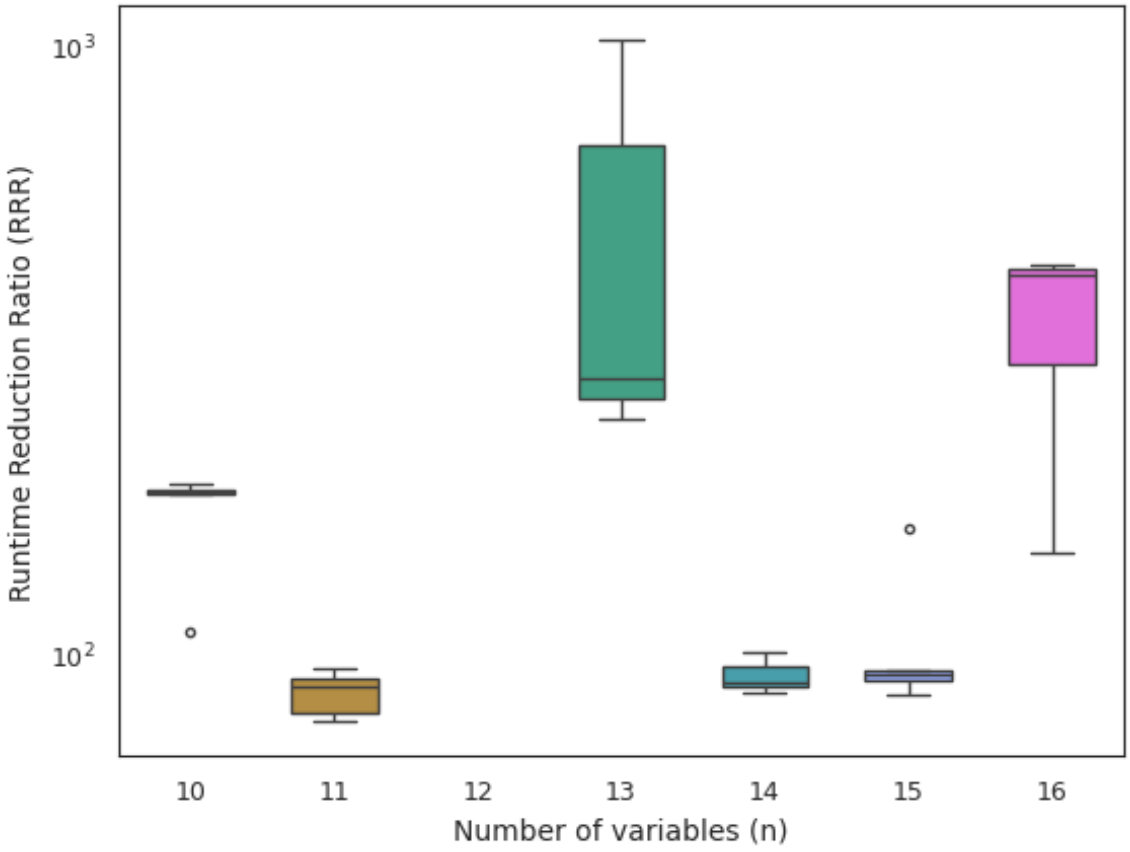

Figure 2: Runtime Reduction Ratio (RRR) for Barabási-Albert graphs.

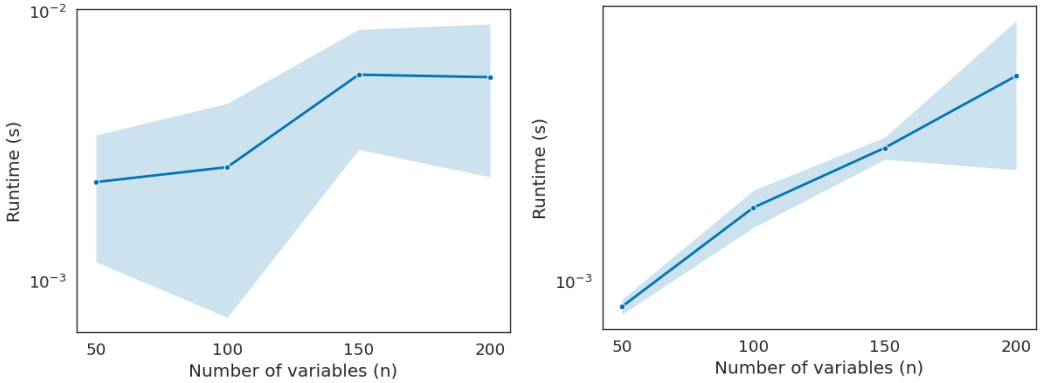

Figure 3: Runtime scalability for large variable counts where exact HP becomes intractable, for Erdős–Rényi (left) and Barabási-Albert (right) graphs.

Figure 4 presents the ISR and robustness results for Erdős–Rényi and Barabási-Albert graphs, respectively, for $n = 10$ and $n = 16$ nodes. The ISR values are close to 1 across a range of λ values, indicating that our method finds interventions of comparable sparsity to the exact HP method. Indeed, it sometimes finds a sparser cause, since the HP causes need not be unique (Halpern, 2016). As λ increases, ISR tends to increase slightly, reflecting a trade-off where larger λ values encourage denser solutions due to the robustness penalty in Equation 6. It is also notable that robustness values are approximately zero for sufficiently large λ, indicating effective violation of the effect.

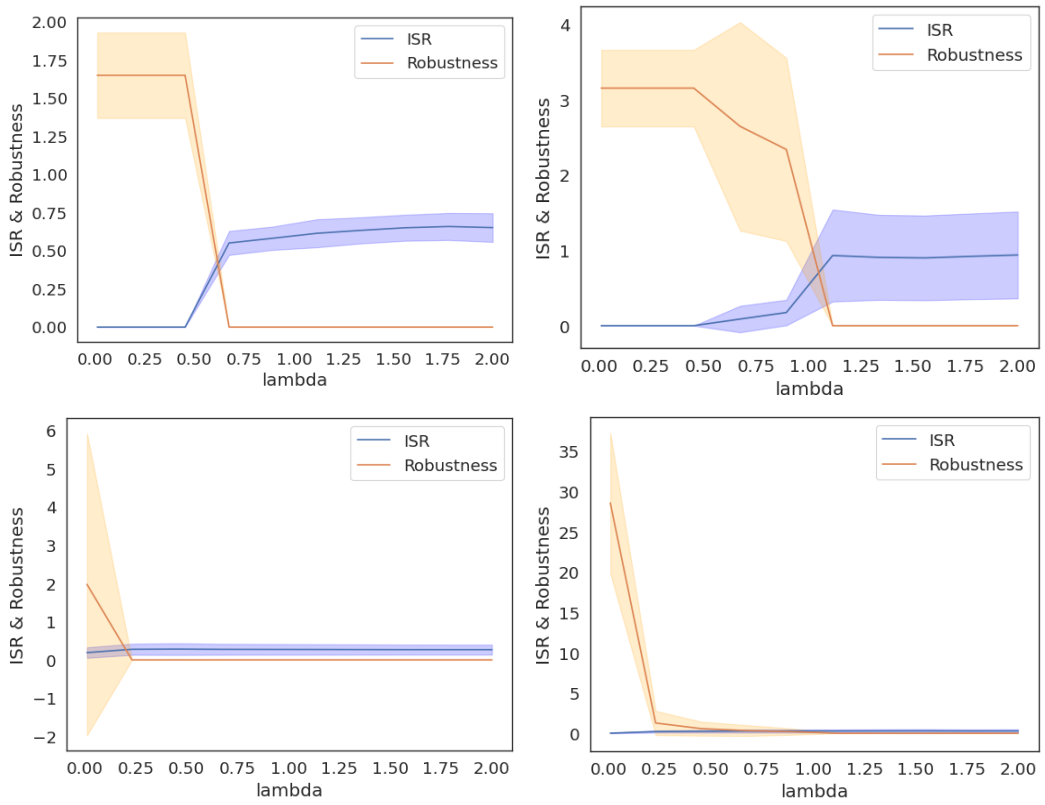

Figure 4: Intervention Sparsity Ratio & robustness vs sparsity parameter λ for Erdős–Rényi graphs (top) and Barabási-Albert graphs (bottom), with $n = 10$ (left) and $n = 16$ (right) nodes. ISR values close to 1 indicate that our method finds interventions of comparable sparsity to the exact HP method. Robustness values near 0 indicate effective violation of the effect.

5. Conclusion

We introduced a differentiable approximation of Halpern–Pearl actual causality for continuous structural causal models. Our approach combines three key innovations: (i) robustness semantics to replace binary formula satisfaction with continuous measures, (ii) additive interventions to replace binary hard interventions with continuous, differentiable ones, and (iii) sparsity relaxation using ℓ_1 penalties to encourage minimal causal explanations while maintaining differentiability.

Empirical evaluation on synthetic Erdős–Rényi and Barabási-Albert graphs with polynomial structural equations demonstrates significant computational advantages over exact HP methods, with runtime improvements that scale favorably as the number of variables increases. Our method achieves intervention sparsity ratios close to 1 and effective robustness violation, indicating faithful approximation of true HP causes while maintaining the essential properties of actual causality.

This work extends the applicability of principled actual causality beyond discrete Boolean systems to continuous, high-dimensional domains. The framework provides tunable control over the sparsity-robustness trade-off via the hyperparameter λ, enabling practitioners to balance computational efficiency with causal fidelity according to their specific requirements. Future work could explore applications to real-world domains such as explainable AI, safety-critical systems, and automated root cause analysis.

References

Derya Aksaray, Austin Jones, Zhaodan Kong, Mac Schwager, and Calin Belta. Q-Learning for robust satisfaction of signal temporal logic specifications. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 6565–6570, December 2016. doi: 10.1109/CDC.2016.7799279. URL https://ieeexplore.ieee.org/abstract/document/7799279.

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, January 2002. ISSN 0034-6861, 1539-0756. doi: 10.1103/RevModPhys.74.47. URL https://link.aps.org/doi/10.1103/RevModPhys.74.47.

Gadi Aleksandrowicz, Hana Chockler, Joseph Y. Halpern, and Alexander Ivrii. The computational complexity of structure-based causality. *Journal of Artificial Intelligence Research*, 58:431–451, 2017. doi: 10.1613/jair.5229.

Sander Beckers and Joost Vennekens. A principled approach to defining actual causation. *Synthese*, 195(2):835–862, February 2018. ISSN 1573-0964. doi: 10.1007/s11229-016-1247-1. URL https://doi.org/10.1007/s11229-016-1247-1.

Sander Beckers, Hana Chockler, and Joseph Y. Halpern. A causal analysis of harm. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/100c1f131893d3b4b34bb8db49bef79f-Paper-Conference.pdf.

Sander Beckers, Hana Chockler, and Joseph Y. Halpern. Quantifying harm. In *Proceedings of Thirty-Second International Joint Conference on Artificial Intelligence*, 2023. doi: 10.24963/ijcai.2023/41.

Ilan Beer, Shoham Ben-David, Hana Chockler, Avigail Orni, and Richard Trefler. Explaining counterexamples using causality. *Formal Methods in System Design*, 40(1):20–40, February 2012. ISSN 1572-8102. doi: 10.1007/s10703-011-0132-2. URL https://doi.org/10.1007/s10703-011-0132-2.

Michel Besserve and Bernhard Schölkopf. Learning soft interventions in complex equilibrium systems. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 170–180. PMLR, August 2022. URL https://proceedings.mlr.press/v180/besserve22a.html. ISSN: 2640-3498.

Gianluca Carloni, Andrea Berti, and Sara Colantonio. The role of causality in explainable artificial intelligence, September 2023. URL http://arxiv.org/abs/2309.09901. arXiv:2309.09901 [cs].

Hana Chockler and Joseph Y. Halpern. Responsibility and blame: a structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004. doi: 10.48550/arXiv.cs/0312038. arXiv:cs/0312038.

Hana Chockler, Daniel Kroening, and Youcheng Sun. Explanations for occluded images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chockler_Explanations_for_Occluded_Images_ICCV_2021_paper.html.

T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to Algorithms, fourth edition*. MIT Press, 2022. ISBN 9780262046305. URL https://books.google.co.uk/books?id=1NKMEAAAQBAJ.

Leonardo de Moura and Nikolaj Bjørner. Z3: an efficient smt solver. In C. R. Ramakrishnan and Jakob Rehof, editors, *Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer, 2008. ISBN 978-3-540-78800-3. doi: 10.1007/978-3-540-78800-3_24.

Alexandre Donzé and Oded Maler. Robust Satisfaction of Temporal Logic over Real-Valued Signals. In Krishnendu Chatterjee and Thomas A. Henzinger, editors, *Formal Modeling and Analysis of Timed Systems*, pages 92–106, Berlin, Heidelberg, 2010. Springer. ISBN 978-3-642-15297-9. doi: 10.1007/978-3-642-15297-9_9.

Clemens Dubslaff, Kallistos Weis, Christel Baier, and Sven Apel. Feature causality. *Journal of Systems and Software*, 209:111915, March 2024. ISSN 0164-1212. doi: 10.1016/j.jss.2023.111915. URL https://www.sciencedirect.com/science/article/pii/S0164121223003102.

Frederick Eberhardt and Richard Scheines. Interventions and Causal Inference. *Philosophy of Science*, 74(5):981–995, December 2007. ISSN 0031-8248, 1539-767X. doi: 10.1086/525638. URL https://www.cambridge.org/core/journals/philosophy-of-science/article/interventions-and-causal-inference/3874FEE8636D10E3F55B2EA46A532006.

P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6(3-4):290–297, July 2022. ISSN 00333883. doi: 10.5486/PMD.1959.6.3-4.12. URL https://publi.math.unideb.hu/load_doi.php?pdoi=10_5486_PMD_1959_6_3_4_12.

Georgios E. Fainekos. *Robustness of Temporal Logic Specifications*. PhD thesis, University of Pennsylvania, 2008. URL https://www.public.asu.edu/~gfaineko/pub/fainekos_thesis.pdf.

Georgios E. Fainekos and George J. Pappas. Robustness of temporal logic specifications for continuous-time signals. *Theoretical Computer Science*, 410(42):4262–4291, September 2009. ISSN 03043975. doi: 10.1016/j.tcs.2009.06.021. URL https://linkinghub.elsevier.com/retrieve/pii/S0304397509004149.

Joseph Y. Halpern. A modification of the Halpern-Pearl definition of causality. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 3022–3033. AAAI Press, July 2015. ISBN 978-1-57735-738-4.

Joseph Y. Halpern. *Actual causality*. MIT Press, 2016. ISBN 978-0-262-03502-6.

Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach — Part 1: Causes, 2001. URL http://arxiv.org/abs/1301.2275. arXiv:1301.2275 [cs].

Joseph Y. Halpern and Judea Pearl. Causes and explanations: a structural-model approach. part i: causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005. ISSN 0007-0882.

Amjad Ibrahim and Alexander Pretschner. From Checking to Inference: Actual Causality Computations as Optimization Problems. volume 12302, pages 343–359. 2020. doi: 10.1007/978-3-030-59152-6_19. URL http://arxiv.org/abs/2006.03363. arXiv:2006.03363 [cs].

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 9551–9561. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6cd9313ed34ef58bad3fdd504355e72c-Abstract.html.

Jhon Mackie. Causes and Conditions. *American Philosophical Quarterly*, 2(4):245–264, 1965. ISSN 0003-0481. URL https://www.jstor.org/stable/20009173. Publisher: [North American Philosophical Publications, University of Illinois Press].

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal Abstraction with Soft Interventions. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, pages 68–87. PMLR, August 2023. URL https://proceedings.mlr.press/v213/massidda23a.html. ISSN: 2640-3498.

Alexandra Meliou, Wolfgang Gatterbauer, Joseph Halpern, Christoph Koch, Katherine Moore, and Dan Suciu. Causality in Databases. *IEEE Data Eng. Bull.*, 33:59–67, January 2010.

Michael Moore. Causation in the law. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition, 2019. URL https://plato.stanford.edu/archives/win2019/entries/causation-law/.

David Fate Norton and Mary J. Norton. *David Hume: A Treatise of Human Nature: Volume 1: Texts*. OUP Oxford, January 2011. ISBN 978-0-19-959633-1. Google-Books-ID: VHCISQAACAAJ.

Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, September 2009. ISBN 978-0-521-89560-6.

Spencer Peters and Joseph Y. Halpern. Causal Modeling With Infinitely Many Variables, December 2021. URL http://arxiv.org/abs/2112.09171. arXiv:2112.09171 [cs].

Hadrien Pouget, Hana Chockler, Youcheng Sun, and Daniel Kroening. Ranking Policy Decisions, October 2021. URL http://arxiv.org/abs/2008.13607. arXiv:2008.13607 [cs, stat].

Marina Sarda Gou, Gabriella Lakatos, Patrick Holthaus, Ben Robins, S Moros, L. Jai Wood, Hugo Da Silva Araujo, C.A.E. deGraft Hanson, Mohammadreza Mousavi, and Farshid Amirabdollahian. Kaspar explains: the effect of causal explanations on visual perspective taking skills in children with autism spectrum disorder. In *Proceedings of the 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2023)*. IEEE, 2023.

Stephen Willard. *General topology*. Courier Corporation, 2012. ISBN 978-0-486-13178-8.

Richard W. Wright. The NESS Account of Natural Causation: A Response to Criticisms. In Richard Goldberg, editor, *Perspectives on Causation*. Hart Publishing, 2011. URL https://ssrn.com/abstract=1918405. Section: 14.

Vicky Jiaqi Zhang, Chandler Squires, and Caroline Uhler. Matching a Desired Causal State via Shift Interventions. November 2021. URL https://openreview.net/forum?id=Sgqb8b8swh7.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning, November 2018. URL http://arxiv.org/abs/1803.01422. arXiv:1803.01422 [cs, stat].