ON THE SUFFICIENCY OF REHEARSAL IN AVOIDING UNDESIRED FUTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

When a model predicts an undesired outcome, deciding how to avoid it naturally calls for identifying causal relations between actions and outcomes. Prior studies suggest that effective actions can often be chosen by judging their *influence* on desired outcomes through *rehearsal* in a hypothetical world, even without faithfully modeling the underlying causal mechanisms. This raises a foundational question: whether rehearsal generally suffices for decision-making without identifying those mechanisms, and if so, what information it requires beyond vanilla prediction. We address this by formalizing the influence information required for decision-making with structural equation models and showing that this information does not in general entail complete causal identification. We further characterize the conditions under which the desired influence is discernible from observational data and develop a principled procedure for determining it under partial structural knowledge. Our theoretical and empirical results support the feasibility of rehearsal in avoiding undesired outcomes with limited structural knowledge.

1 INTRODUCTION

Machine learning models have proven successful as predictors: when trained on observational data, they can accurately predict outcomes for unseen instances (Vapnik, 1995; Bishop, 2006; Goodfellow et al., 2016). Prediction alone, however, is insufficient when the predicted outcomes are undesired: it warns us what is likely to happen, but not what to do about it. Acting on such a warning requires identifying which actions can affect the undesired outcomes. This challenge, known as the *avoiding undesired future* (AUF) problem, calls for a shift from prediction to decision (Zhou, 2022).

Confronting the AUF problem requires models with decision-supporting capabilities that go beyond statistical predictions. Causal models (Pearl, 2009; Peters et al., 2017) are helpful, while a comprehensive characterization of environments is unnecessary and not always available in AUF scenarios. What matters in AUF is whether actions can shape a more desired future, rather than fully explaining why future outcomes may arise; e.g., humans often make good decisions without a thorough or faithful understanding of the surrounding environment. Moreover, the goal of uncovering the causal mechanisms underlying historical data is inherently challenging (Chickering et al., 1995), and even if uncovered, they may no longer hold by the time actions are taken in dynamic environments (Zhou, 2022). To this end, (Zhou, 2022; 2023) the concept of *influence* to facilitate learning to discern which actions can shape desired outcomes, requiring models to be more informative than statistical associations for prediction and less demanding than causal identification. Hence, a series of *rehearsal learning* techniques have been developed to learn to rehearse potential actions and seek those with the desired influence on future outcomes, and have been shown to be empirically promising in various AUF tasks (Qin et al., 2023; 2025; Du et al., 2024; 2025a;b; Tao et al., 2026b).

Given the recent advances in rehearsal learning techniques, a foundational question remains open: *how does rehearsal learning theoretically relate to statistical and causal learning?* Intuitively, since their learning goals are quite different, a model adequate for one goal may be inadequate for another. Is this intuition always true, or are there cases where a model learned for one suffices for another? To answer this question, we need to characterize what information rehearsal requires, to what extent it goes beyond statistical prediction, and whether it reaches as far as causal identification.

In this paper, we present a formal framework that relates the information requirements of rehearsal, statistical, and causal learning. Under the semantics of structural causal models, we characterize the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

influence information required by decision-making for AUF and compare it with valuations at the statistical and causal levels. We show that the interventional effects pursued by causal identification always entail the influence indicators for rehearsing decisions, whereas the converse does *not* hold: for any underlying data-generating model, there exists a different model that agrees with it on every influence indicator yet disagrees on interventional effects (Theorem 1). Thus, decision-making for AUF generally requires less information than identifying the interventional effects. We also show that the statistical information for prediction does not suffice for AUF, and the influence information does not suffice for identifying higher levels of interventional and counterfactual effects (Corollary 1). These results suggest that, while statistical models fall short of supporting decision-making in the AUF problem, rehearsal learning generally suffices without relying on causal identification.

Further, to study the extent to which the structural knowledge required for AUF goes beyond statistical prediction, we characterize the conditions under which the influence of actions on desired outcomes is discernible from observational data. We show that knowing the full graph structure or the common causes of actionable variables and context or outcome variables is sufficient for discernibility (Propositions 1 and 2). We also develop a sound and complete procedure for determining discernibility under partial knowledge of the common causes (Algorithm 1 and Theorem 2). Moreover, for practical scenarios where structural knowledge is limited, we provide a novel and cost-effective approach to elicit partial knowledge on demand from domain experts (Algorithm 2), and demonstrate its efficacy in synthetic and real-world applications. In summary, our theoretical and experimental results both indicate that the AUF problem can be addressed through rehearsal learning with substantially weaker structural knowledge than that required for causal identification.

2 PRELIMINARIES

We represent random variables with capital letters (X) and their realized values with lowercase letters (x), and use boldface for sets of variables ($\mathbf{X}$) and their joint values ($\mathbf{x}$). For a variable X, Ω_X denotes its domain, and the joint domain of a set $\mathbf{X}$ is the Cartesian product s $\Omega_{\mathbf{X}} := \prod_{X \in \mathbf{X}} \Omega_X$.

The AUF problem. The AUF problem concerns a system of random variables $\mathbf{V}$, in which an agent can observe some context variables and act on others to influence outcome variables $\mathbf{Y} \subseteq \mathbf{V}$. The desirability of these outcomes is represented by a desired set $\mathcal{S} \subseteq \Omega_{\mathbf{Y}}$. Throughout this work, $\mathbf{Y}$ and $\mathcal{S}$ are pre-specified, with $\emptyset \subsetneq \mathbf{Y} \subsetneq \mathbf{V}$ and $\emptyset \subsetneq \mathcal{S} \subsetneq \Omega_{\mathbf{Y}}$. An AUF warning is issued if, after observing $\mathbf{X} = \mathbf{x}$ with $P(\mathbf{X} = \mathbf{x}) > 0$, the undesired event is likely to happen with higher probability, i.e., $P(\mathbf{Y} \notin \mathcal{S} | \mathbf{X} = \mathbf{x}) > P(\mathbf{Y} \in \mathcal{S} | \mathbf{X} = \mathbf{x})$ (Zhou, 2022). Hence, the goal is to choose an action that is anticipated to avoid the undesired event indicated by the warning. Resolving the warning requires the ability to *influence* future outcomes by acting on variables in $\mathbf{V}$. We next introduce the model semantics used to formalize these actions and their hypothetical outcomes.

Structural causal models. We use the language of *structural causal models* (Pearl, 2009) to formalize potential actions and outcomes. We consider *recursive* models with finite discrete domains.

Definition 1 (Structural Causal Model (SCM))**.** An SCM is a tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where $\mathbf{U}$ is a set of noise variables determined by factors outside the model, $\mathbf{V}$ is a set of endogenous variables, $\mathcal{F} = \{f_V\}_{V \in \mathbf{V}}$ is a set of structural functions, and $P(\mathbf{U})$ is a probability distribution over the domain of $\mathbf{U}$. Each f_V maps the values of parents $\mathbf{Pa}_V \subseteq \mathbf{V} \setminus \{V\}$ and noise $U_V \in \mathbf{U}$ to V, written as $V \leftarrow f_V(\mathbf{pa}_V, u_V)$. The entire set $\mathcal{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$, which uniquely determines the values of all variables $\mathbf{V}(\mathbf{u})$ given a realization of the noises $\mathbf{u}$.

An SCM $\mathcal{M}$ induces a causal graph $\mathcal{G}$ over $\mathbf{V}$, where a directed edge $V_j \to V_i$ exists if $V_j \in \mathbf{Pa}_{V_i}$. Since $\mathcal{M}$ is recursive, $\mathcal{G}$ is acyclic. For a set $\mathbf{W} \subseteq \mathbf{V}$, we write $\mathrm{Anc}_{\mathcal{G}}(\mathbf{W}) \subseteq \mathbf{V} \setminus \mathbf{W}$ for the set of ancestors of $\mathbf{W}$ in $\mathcal{G}$, i.e., the variables having a directed path in $\mathcal{G}$ to some variable in $\mathbf{W}$. We write $\mathcal{G}_{\underline{\mathbf{W}}}$ for the subgraph obtained by deleting all edges emerging from variables in $\mathbf{W}$. For two disjoint sets $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{V}$, we write $\mathbf{X} \prec \mathbf{Z}$ when no variable in $\mathbf{Z}$ is an ancestor of any variable in $\mathbf{X}$, i.e., $(\mathbf{X} \cup \mathrm{Anc}_{\mathcal{G}}(\mathbf{X})) \cap \mathbf{Z} = \emptyset$. In the AUF problem, given the contexts $\mathbf{X}$ and outcomes $\mathbf{Y}$, we require the actionable variables $\mathbf{Z}$ to satisfy the admissibility condition $\mathbf{X} \prec \mathbf{Z} \prec \mathbf{Y}$. This reflects the forward-looking nature of rehearsal: an action cannot change an already observed context and is taken before the outcomes it seeks to influence are realized.

The mechanisms $\mathcal{F}$ in an SCM $\mathcal{M}$ remain invariant unless explicitly altered. Evaluating them at a realization $\mathbf{u}$ yields the unique solution $\mathbf{V}(\mathbf{u})$. Altering $\mathbf{Z} \subseteq \mathbf{V}$ to $\mathbf{z} \in \Omega_{\mathbf{Z}}$ replaces the mechanisms

of $\mathbf{Z}$ with the constants $\mathbf{z}$. The resulting solution $\mathbf{V}_{\mathbf{z}}(\mathbf{u})$ is obtained by evaluating $\mathcal{F}_{\mathbf{z}} := \{f_V : V \in \mathbf{V} \setminus \mathbf{Z}\} \cup \{f_V \leftarrow z_V : V \in \mathbf{Z}\}$ at $\mathbf{u}$. For any $\mathbf{W} \subseteq \mathbf{V}$, the corresponding solution $\mathbf{W}_{\mathbf{z}}(\mathbf{u})$ is called the potential outcome of $\mathbf{W}$ under the alteration $\mathbf{z}$. When applied in reality, this mechanism replacement is formally represented by the do-operator $do(\mathbf{Z} = \mathbf{z})$ (Pearl, 2009; Peters et al., 2017).

Observational and interventional valuations. An SCM induces valuations at multiple levels of the causal hierarchy (Shpitser & Pearl, 2008; Pearl & Mackenzie, 2018; Bareinboim et al., 2022). We first recall the observational and interventional levels, which range over pairs of sets of variables.

Definition 2 (Observational Valuation)**.** An SCM $\mathcal{M}$ induces $\mathcal{L}_O(\mathcal{M})$, a collection of probability mass functions $P^{\mathcal{M}}(\mathbf{Z}|\mathbf{X})$ for all pairs of sets $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{V}$ with $\mathbf{X} \cap \mathbf{Z} = \emptyset$. For each realization $\mathbf{x} \in \Omega_{\mathbf{X}}$ and $\mathbf{z} \in \Omega_{\mathbf{Z}}$ with $\sum_{\{\mathbf{u}:\, \mathbf{X}(\mathbf{u})=\mathbf{x}\}} P(\mathbf{u}) > 0$, the probability is given by

$$P^{\mathcal{M}}(\mathbf{Z} = \mathbf{z} \mid \mathbf{X} = \mathbf{x}) := \textstyle\sum_{\{\mathbf{u}:\, \mathbf{X}(\mathbf{u})=\mathbf{x},\, \mathbf{Z}(\mathbf{u})=\mathbf{z}\}} P(\mathbf{u}) \big/ \sum_{\{\mathbf{u}:\, \mathbf{X}(\mathbf{u})=\mathbf{x}\}} P(\mathbf{u}).$$

The observational valuation is standard in statistical learning, which usually aims to estimate the conditional distribution $P(\mathbf{Z}|\mathbf{X})$ from data, predicting the target variables $\mathbf{Z}$ given the features $\mathbf{X}$. In contrast, the goal of causal learning is to identify and estimate the interventional distribution $P(\mathbf{Z}|do(\mathbf{X}))$ (a.k.a. causal effect) that characterizes how variables $\mathbf{Z}$ respond to interventions on $\mathbf{X}$.

Definition 3 (Interventional Valuation)**.** An SCM $\mathcal{M}$ induces $\mathcal{L}_I(\mathcal{M})$, a collection of probability mass functions $P^{\mathcal{M}}(\mathbf{Z}|do(\mathbf{X}))$ for all pairs of sets $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{V}$ with $\mathbf{X} \cap \mathbf{Z} = \emptyset$. For each assignment $\mathbf{x} \in \Omega_{\mathbf{X}}$ and realization $\mathbf{z} \in \Omega_{\mathbf{Z}}$, the probability is given by

$$P^{\mathcal{M}}(\mathbf{Z} = \mathbf{z} \mid do(\mathbf{X} = \mathbf{x})) := \textstyle\sum_{\{\mathbf{u}:\, \mathbf{Z}_{\mathbf{x}}(\mathbf{u})=\mathbf{z}\}} P(\mathbf{u}).$$

In contrast to quantities at the observational level, which can be reliably estimated from observational data given sufficient samples, quantities at the interventional level are generally *not* uniquely determined by observational data alone. Specifically, two distinct SCMs $\mathcal{M}_1, \mathcal{M}_2$ may satisfy $\mathcal{L}_O(\mathcal{M}_1) = \mathcal{L}_O(\mathcal{M}_2)$ while inducing different interventional distributions $\mathcal{L}_I(\mathcal{M}_1) \neq \mathcal{L}_I(\mathcal{M}_2)$. Consequently, identifying interventional distributions generally requires information beyond observational data, such as the structural knowledge represented by the causal graph $\mathcal{G}$ (Pearl, 2009).

Beyond the interventional level, the causal hierarchy extends to even higher levels involving counterfactual quantities (Shpitser & Pearl, 2007; Raghavan & Bareinboim, 2025), e.g., the joint statements about potential outcomes $\mathbf{Y}_{\mathbf{z}}(\mathbf{u})$ and $\mathbf{Y}_{\mathbf{z}'}(\mathbf{u})$ under conflicting interventions $\mathbf{z} \neq \mathbf{z}'$, evaluated at the same realization $\mathbf{u}$ of the noises. The resulting counterfactual valuation induced by $\mathcal{M}$ is denoted by $\mathcal{L}_C(\mathcal{M})$; see Definition 10 in Appendix A.2 for a formal definition.

Rehearsal. To address the AUF problem, *rehearsal* has been introduced as a forward-looking process for discerning the influence of actionable variables on desired outcomes (Zhou, 2022; 2023). This process evaluates, as a thought experiment in a hypothetical world, the desirability of future outcomes that an action would bring about. An SCM $\mathcal{M}$ employed to evaluate hypothetical outcomes is called a rehearsal model, which need not be the true SCM $\mathcal{M}^*$ underlying the AUF task. Formally, given context variables $\mathbf{X} \subseteq \mathbf{V} \setminus \mathbf{Y}$, we consider alterations on some actionable variables $\mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ satisfying $\mathbf{X} \prec \mathbf{Z} \prec \mathbf{Y}$. Let $\Delta_{\mathbf{Z}} := \prod_{Z \in \mathbf{Z}} \Delta_Z$ denote the action space, where $\Delta_Z := \Omega_Z \cup \{\perp\}$ and $\perp$ denotes no alteration. An action $\mathbf{z} \in \Delta_{\mathbf{Z}}$ specifies the altered subset of variables $\mathbf{Z}' := \{Z \in \mathbf{Z} : z_Z \neq \perp\}$ with the alteration values $\mathbf{z}' \in \Omega_{\mathbf{Z}'}$ (i.e., the projection of $\mathbf{z}$ onto $\mathbf{Z}'$). With a slight abuse of notation, we set $\mathbf{Y}_{\mathbf{z}}(\mathbf{u}) := \mathbf{Y}_{\mathbf{z}'}(\mathbf{u})$; when every component of $\mathbf{z}$ is $\perp$, no alteration is applied and $\mathbf{Y}_{\mathbf{z}}(\mathbf{u}) = \mathbf{Y}(\mathbf{u})$. Then, given context $\mathbf{X} = \mathbf{x}$, the belief about the future outcomes under action $\mathbf{z}$ is represented by the rehearsal probability $P^{\mathcal{M}}(\mathbf{Y}_{\mathbf{z}}|\mathbf{X} = \mathbf{x})$. Accordingly, the rehearsal process determines whether action $\mathbf{z}$ resolves the AUF warning by evaluating $P^{\mathcal{M}}(\mathbf{Y}_{\mathbf{z}} \in \mathcal{S}|\mathbf{X} = \mathbf{x}) \geq P^{\mathcal{M}}(\mathbf{Y}_{\mathbf{z}} \notin \mathcal{S}|\mathbf{X} = \mathbf{x})$. In this way, rehearsal uses the warning of an undesired event as a basis for shaping future outcomes toward desired ones.

3 The Rehearsal Level

In the AUF problem, the knowledge about the world required for rehearsal goes beyond the observational level. To characterize how far beyond it reaches, in this section we first give a formal definition of the rehearsal level and then analyze its relationship with the interventional level. The rehearsal valuation is given as follows, in the same spirit as the observational and interventional valuations.

Definition 4 (Rehearsal Valuation)**.** An SCM $\mathcal{M}$ induces $\mathcal{L}_R(\mathcal{M})$, a collection of influence indicators $I^{\mathcal{M}}(\mathbf{Z}|\mathbf{X})$ for all pairs of sets of variables $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{V} \setminus \mathbf{Y}$ with $\mathbf{X} \cap \mathbf{Z} = \emptyset$. For each realization $\mathbf{x} \in \Omega_{\mathbf{X}}$ and action $\mathbf{z} \in \Delta_{\mathbf{Z}}$ with $\sum_{\{\mathbf{u}:\, \mathbf{X}_{\mathbf{z}}(\mathbf{u})=\mathbf{x}\}} P(\mathbf{u}) > 0$, the indicator is given by:

$$I^{\mathcal{M}}(\mathbf{Z}=\mathbf{z} \mid \mathbf{X}=\mathbf{x}) \coloneqq \mathbb{1}\left[\sum_{\{\mathbf{u}:\, \mathbf{X}_{\mathbf{z}}(\mathbf{u})=\mathbf{x}, \mathbf{Y}_{\mathbf{z}}(\mathbf{u}) \in \mathcal{S}\}} P(\mathbf{u}) \geq \sum_{\{\mathbf{u}:\, \mathbf{X}_{\mathbf{z}}(\mathbf{u})=\mathbf{x}, \mathbf{Y}_{\mathbf{z}}(\mathbf{u}) \notin \mathcal{S}\}} P(\mathbf{u})\right].$$

For an AUF task satisfying $\mathbf{X} \prec \mathbf{Z} \prec \mathbf{Y}$, $I^{\mathcal{M}}(\mathbf{Z}=\mathbf{z}|\mathbf{X}=\mathbf{x}) = 1$ indicates that, given the observed context $\mathbf{x}$, the action $\mathbf{z}$ shifts the odds in favor of the desired event, making it at least as probable as the undesired event. This resolves the AUF warning issued from observation.

Remark 1. Definition 4 ranges over all disjoint pairs of variable sets $\mathbf{X}, \mathbf{Z}$ for consistency with Definitions 2 and 3, with the desired outcomes $\mathbf{Y}$ pre-specified. It evaluates $\mathbf{X}_{\mathbf{z}}$ and $\mathbf{Y}_{\mathbf{z}}$ under the same alteration. For variables satisfying $\mathbf{X} \prec \mathbf{Z} \prec \mathbf{Y}$, which is the case in forward-looking AUF tasks as described in Section 2, we have $\mathbf{X}_{\mathbf{z}}(\mathbf{u}) = \mathbf{X}(\mathbf{u})$. That is, variables in $\mathbf{X}$ are not affected or altered by the action $\mathbf{z}$. In this sense, rehearsal differs from counterfactual reasoning (Pearl, 2009): the counterfactual imagines the past (i.e., what *would have* followed from an action not taken), whereas rehearsal imagines the future (i.e., what *will* follow from an action not yet taken).

Remark 2. Given an SCM $\mathcal{M}$, the influence indicator actually encodes the inequality $P^{\mathcal{M}}(\mathbf{Y}_{\mathbf{z}} \in \mathcal{S}|\mathbf{X})/P^{\mathcal{M}}(\mathbf{Y}_{\mathbf{z}} \notin \mathcal{S}|\mathbf{X}) \geq 1$. This is equivalent to requiring that the *odds* of desired outcomes relative to undesired outcomes be at least one. Because probabilities represent the agent's subjective degrees of belief about the future, what matters is the relative strength rather than the absolute value. Thus, we view the odds of one as a natural threshold for the influence indicator, reflecting the intuitive principle that the more probable outcome prevails. Any other positive threshold can be similarly adopted and accommodated in the same way in the following analysis.

To compare the information encoded at each level of the hierarchy, we formalize when two SCMs are indistinguishable at a given level. Here, the subscript i in $\mathcal{L}_i$ denotes the level, which can be O, R, I, or C for the observational, rehearsal, interventional, and counterfactual levels, respectively.

Definition 5 ($\mathcal{L}_i$-Consistency)**.** For any two SCMs $\mathcal{M}, \mathcal{M}'$ and a level $i \in \{O, R, I, C\}$, $\mathcal{M}$ is *$\mathcal{L}_i$-consistent* with $\mathcal{M}'$, denoted as $\mathcal{M} \sim_i \mathcal{M}'$, if the $\mathcal{L}_i$ induced by $\mathcal{M}$ equals that induced by $\mathcal{M}'$.

Based on Definition 5, we ask whether agreement at one level implies agreement at another, i.e., whether one level's statements fully determine another's. Denote by $\mathbb{M}$ the set of recursive SCMs.

Definition 6 (Implication Relative to $\mathcal{M}^*$)**.** For a level $i \in \{O, R, I, C\}$, $\mathcal{L}_i$ is said to *imply* $\mathcal{L}_j$ relative to $\mathcal{M}^*$ if $\mathcal{M} \sim_i \mathcal{M}^*$ implies $\mathcal{M} \sim_j \mathcal{M}^*$ for all $\mathcal{M} \in \mathbb{M}$.

The following theorem characterizes the relationship between the rehearsal and interventional levels.

Theorem 1. *For any SCM $\mathcal{M}^* \in \mathbb{M}$, $\mathcal{L}_R$ does not imply $\mathcal{L}_I$ relative to $\mathcal{M}^*$, while the converse holds, i.e., $\mathcal{L}_I$ implies $\mathcal{L}_R$ relative to $\mathcal{M}^*$.*

The proof is provided in Appendix A.1. Theorem 1 shows that the interventional level determines the rehearsal level, but not conversely, for every underlying SCM $\mathcal{M}^*$ and every fixed choice of $\mathbf{Y}$ and $\mathcal{S}$ satisfying $\emptyset \subsetneq \mathbf{Y} \subsetneq \mathbf{V}$ and $\emptyset \subsetneq \mathcal{S} \subsetneq \Omega_{\mathbf{Y}}$. The rehearsal valuation is thus a task-specific coarsening of interventional information, retaining binary influence judgments rather than exact probabilities. In other words, decision-making for AUF does not generally require causal identification: an agent can determine whether an action resolves the AUF warning without completely identifying the underlying causal effects. Figure 1 intuitively depicts this result: the SCMs $\mathcal{L}_I$-consistent with $\mathcal{M}^*$ form a proper subset of those $\mathcal{L}_R$-consistent with $\mathcal{M}^*$. Rehearsal is therefore, in principle, informationally weaker than causal identification. An example of two SCMs that agree at the observational and rehearsal levels yet disagree at the interventional level is given in Example 1 of Appendix C.1.

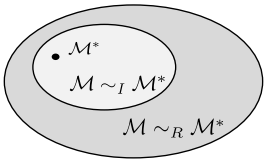

Figure 1: $\mathcal{L}_I$ implies $\mathcal{L}_R$, but not vice versa.

We further compare the rehearsal level with the observational and counterfactual levels as follows.

Corollary 1. *Assume that $|\Omega_{\mathbf{V}\setminus\mathbf{Y}}| \geq 3$. Let $\mathcal{I} := \{O, R, I, C\}$, ordered as $O < R < I < C$. Then, for all $i, j \in \mathcal{I}$ satisfying $i < j$, with respect to the Lebesgue measure over (a suitable encoding of $\mathcal{L}_C$-equivalence classes of) SCMs, the subset in which $\mathcal{L}_i$ implies $\mathcal{L}_j$ has measure zero.*

The proof is given in Appendix A.1. The cardinality condition is mild, e.g., it holds with one ternary actionable variable, which provides enough freedom for observationally equivalent models to disagree across the odds of one. Corollary 1 yields a separation in two directions. On the one side, observational information does not determine the rehearsal level. While the AUF warning is triggered at the observational level, the action that resolves it is not determined at that level. On the other side, rehearsal information does not determine the two causal levels: it leaves the causal effects at the interventional level unidentified and, beyond that, cannot answer counterfactual questions about what would have followed from actions that were not taken. An additional discussion on the relationship between the observational and rehearsal levels is provided in Appendix A.6.

4 Influence Discernibility

As shown in Section 3, observational information alone does not generally determine which action resolves the AUF warning, whereas the interventional level provides more information than the influence indicator requires. This motivates an appeal to Vapnik's principle (Vapnik, 1995): "when solving a given problem, try to avoid solving a more general problem as an intermediate step". Applied to the AUF problem, this principle suggests discerning the desired influence directly, rather than first recovering interventional probabilities and then deriving the influence indicator from them.

Motivated by this principle, we ask whether the desired influence can be discerned from observational data given limited structural information about the underlying SCM.

Definition 7 (Discernibility)**.** Consider an arbitrary SCM $\mathcal{M}^*$ from a class of SCMs $\mathbb{M}^*$ and the observational distribution $P(\mathbf{V})$ induced by $\mathcal{M}^*$. For any $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{V} \setminus \mathbf{Y}$ with $\mathbf{X} \prec \mathbf{Z} \prec \mathbf{Y}$, the influence indicator $I(\mathbf{Z}|\mathbf{X})$ is said to be *discernible* from $\mathbb{M}^*$ and $P(\mathbf{V})$ if and only if $I^{\mathcal{M}_1}(\mathbf{Z}|\mathbf{X}) = I^{\mathcal{M}_2}(\mathbf{Z}|\mathbf{X})$ holds for every pair of SCMs $\mathcal{M}_1, \mathcal{M}_2 \in \mathbb{M}^*$ with $P^{\mathcal{M}_1}(\mathbf{V}) = P^{\mathcal{M}_2}(\mathbf{V}) = P(\mathbf{V}) > 0$.

When $I(\mathbf{Z}|\mathbf{X})$ is discernible, the influence judgments are determined for all contexts and actions. This notion is closely related to partial identification (Balke & Pearl, 1994): the entire indicator can be determined even when the corresponding interventional probabilities are not uniquely identified. The influence indicator enables a direct comparison with causal-effect identification. The resulting influence judgments also support action selection under varying costs or feasibility constraints. We characterize when such discernibility can be established under partial causal knowledge.

For simplicity, in this section, we consider a single actionable variable Z and restrict attention to *Markovian* SCMs, whose variable-specific noises are mutually independent. The partial knowledge studied here concerns common causes among observed variables. Extending to semi-Markovian models would require accounting for latent confounding, which is beyond the scope of this section.

4.1 Sufficient conditions for influence discernibility

In this subsection, we investigate what structural information, together with the observational distribution, is sufficient for determining the influence indicator in Definition 7.

Discernibility from graph structure. Given structural knowledge $\mathcal{K}$ about the true SCM $\mathcal{M}^*$, an SCM is said to be $\mathcal{K}$*-consistent* if it satisfies the statements in $\mathcal{K}$. A typical form of such knowledge is the directed acyclic graph $\mathcal{G}$, which specifies the parents of each endogenous variable. Let $\mathbb{M}_{\mathcal{G}}$ denote the class of SCMs that are Markovian and $\mathcal{G}$-consistent.

Proposition 1. *For any $\mathbf{X} \subseteq \mathbf{V} \setminus \mathbf{Y}$ and $Z \in \mathbf{V} \setminus \mathbf{Y}$ satisfying $\mathbf{X} \prec Z \prec \mathbf{Y}$, given the causal graph $\mathcal{G}$ over $\mathbf{V}$, the influence indicator $I(\mathbf{Z}|\mathbf{X})$ is discernible from $\mathbb{M}_{\mathcal{G}}$ and $P(\mathbf{V})$.*

The proof is provided in Appendix A.3. Proposition 1 shows that knowledge of the causal graph is sufficient for discerning the influence indicator. Under the stated assumptions, such knowledge identifies causal effects (Pearl, 2009, Corollary 3.2.6), and hence also determines the relevant rehearsal probabilities and influence indicators.

Discernibility from common causes. A complete causal graph usually provides more structural information than a particular goal requires, e.g., it is often enough to know the causes relevant to the action and its outcomes for confounding control (Pearl, 2009, Section 3.3). For the AUF problem, contexts $\mathbf{X}$ should also be considered, as conditioning on $\mathbf{X}$ may open paths between Z and $\mathbf{Y}$ that are otherwise blocked. We thus introduce the following common-cause structure.

Definition 8 (Common-Cause Structure)**.** Let $\mathcal{G}$ denote the underlying causal graph induced by the true SCM. For any $\mathbf{X} \subseteq \mathbf{V} \setminus \mathbf{Y}$ and $Z \in \mathbf{V} \setminus \mathbf{Y}$ with $\mathbf{X} \prec Z \prec \mathbf{Y}$, the *common-cause structure* w.r.t. $(\mathbf{X}, Z, \mathbf{Y})$ is defined as $\mathcal{C} := \mathrm{Anc}_{\mathcal{G}_{\underline{Z}}}(Z) \cap \big(\mathrm{Anc}_{\mathcal{G}_{\underline{Z}}}(\mathbf{Y}) \cup \mathrm{Anc}_{\mathcal{G}_{\underline{Z}}}(\mathbf{X})\big)$.

Thus, $\mathcal{C}$ includes both the causes shared by Z and the outcomes $\mathbf{Y}$ and those shared by Z and the contexts $\mathbf{X}$. It records which variables are relevant common causes, without specifying their parent sets or the edges among them. An SCM is $\mathcal{C}$*-consistent* if the common-cause structure it induces for $(\mathbf{X}, Z, \mathbf{Y})$ coincides with $\mathcal{C}$. Let $\mathbb{M}_{\mathcal{C}}$ denote the class of $\mathcal{C}$-consistent SCMs.

Proposition 2. *For any $\mathbf{X} \subseteq \mathbf{V} \setminus \mathbf{Y}$ and $Z \in \mathbf{V} \setminus \mathbf{Y}$ satisfying $\mathbf{X} \prec Z \prec \mathbf{Y}$, given the common-cause structure $\mathcal{C}$ for $(\mathbf{X}, Z, \mathbf{Y})$, the influence indicator $I(Z|\mathbf{X})$ is discernible from $\mathbb{M}_{\mathcal{C}}$ and $P(\mathbf{V})$.*

The proof is provided in Appendix A.4. Proposition 2 shows that the common-cause structure, which is weaker than the causal graph and does not determine the full interventional valuation, is sufficient for discerning the influence indicator. Yet, the structure is able to determine the causal effect relevant to the decision, namely $P(\mathbf{Y}|\mathbf{X}, do(Z))$. Discernibility in this case is therefore obtained with causal identification: the indicator is settled because the underlying interventional effect is identified.

Discernibility without fully knowing the causes. We thus ask whether the knowledge demanded by Proposition 2 can be relaxed further: can the influence indicator be discerned even when the relevant causal effect is not identified? This question is practical, as relevant common causes may not be known exhaustively. For instance, a physician may know that a genetic marker is a common cause of treatment and recovery, while remaining unsure whether an observed comorbidity is also relevant. To represent such incomplete knowledge, we define the following structure.

Definition 9 (Partial-Cause Structure)**.** Let $\mathcal{G}$ denote the underlying causal graph induced by the true SCM. For any $\mathbf{X} \subseteq \mathbf{V} \setminus \mathbf{Y}$ and $Z \in \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ with $\mathbf{X} \prec Z \prec \mathbf{Y}$, a set $\widehat{\mathcal{C}}$ is called a *partial-cause structure* w.r.t. $(\mathbf{X}, Z, \mathbf{Y})$ if $\widehat{\mathcal{C}} \subseteq \mathrm{Anc}_{\mathcal{G}_{\underline{Z}}}(Z) \cap \big(\mathrm{Anc}_{\mathcal{G}_{\underline{Z}}}(\mathbf{Y}) \cup \mathrm{Anc}_{\mathcal{G}_{\underline{Z}}}(\mathbf{X})\big)$.

A partial-cause structure $\widehat{\mathcal{C}}$ specifies that every variable in $\widehat{\mathcal{C}}$ belongs to the common-cause structure (see Figure 2 for an example), while remaining agnostic about the others. To reason under this partial specification, we consider candidate common-cause structures that are compatible with the partial causes.

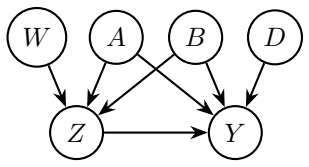

Figure 2: $\widehat{\mathcal{C}} = \{A\}$ is a partial-cause structure w.r.t. $(\emptyset, Z, Y)$, while A and B are the common causes of Z and Y.

Formally, given a partial-cause structure $\widehat{\mathcal{C}}$ w.r.t. $(\mathbf{X}, Z, \mathbf{Y})$, a set $\mathcal{C}'$ is called a *candidate* common-cause structure if $\widehat{\mathcal{C}} \subseteq \mathcal{C}' \subseteq \mathbf{V} \setminus (\{Z\} \cup \mathbf{Y})$. Denote by $\mathbb{C}(\widehat{\mathcal{C}})$ the set of all candidate common-cause structures of $\widehat{\mathcal{C}}$. Then, the set of SCMs that are consistent with $\widehat{\mathcal{C}}$ is expressed as $\mathbb{M}_{\widehat{\mathcal{C}}} = \bigcup_{\mathcal{C} \in \mathbb{C}(\widehat{\mathcal{C}})} \mathbb{M}_{\mathcal{C}}$. Let $\mathbb{C}_P(\widehat{\mathcal{C}}) := \big\{\mathcal{C} \in \mathbb{C}(\widehat{\mathcal{C}}) : \exists\, \mathcal{M} \in \mathbb{M}_{\mathcal{C}},\ P^{\mathcal{M}}(\mathbf{V}) = P(\mathbf{V})\big\}$. Hence, $\mathbb{M}_{\widehat{\mathcal{C}}}$ includes the true SCM $\mathcal{M}^*$, and $\mathbb{C}_P(\widehat{\mathcal{C}})$ includes the common-cause structure $\mathcal{C}$ induced by the true SCM and observational distribution.

For each candidate $\mathcal{C}$, let $I^{\mathcal{C}}(Z|\mathbf{X})$ denote the indicator obtained by applying the adjustment expression in Appendix B to $P(\mathbf{V})$. By Proposition 2, this equals $I^{\mathcal{M}^*}(Z|\mathbf{X})$ when $\mathcal{C}$ is the true common-cause structure. Hence, agreement over a candidate set containing the true structure determines the true indicator. Algorithm 1 checks such agreement over a supplied candidate set $\mathbb{C}$, with the minimum and maximum taken pointwise over contexts and actions. This procedure primarily serves the theoretical characterization of discernibility given below. Its running time is $O(|\mathbb{C}|T_{\mathrm{eval}})$, where T_{eval} denotes the cost of evaluating one candidate indicator. While linear in the number of supplied candidates, this cost can be exponential in the number of unresolved membership choices. Improving computational scalability to larger candidate spaces, e.g., via structure-aware pruning, is left for future work. For the candidate set $\mathbb{C}_P(\widehat{\mathcal{C}})$, the following theorem gives a necessary and sufficient characterization of discernibility.

Algorithm 1 Discerning with candidates

Input: context $\mathbf{X}$; actionable variable Z; $\mathcal{L}_O$ data $P(\mathbf{V})$; candidate set $\mathbb{C}$.
Output: The common indicator over $\mathbb{C}$ if all candidates agree, else `FAIL`.

$I_{\min}(Z|\mathbf{X}) \leftarrow \min_{\mathcal{C} \in \mathbb{C}} I^{\mathcal{C}}(Z|\mathbf{X})$
$I_{\max}(Z|\mathbf{X}) \leftarrow \max_{\mathcal{C} \in \mathbb{C}} I^{\mathcal{C}}(Z|\mathbf{X})$
if $I_{\min}(Z|\mathbf{X}) \neq I_{\max}(Z|\mathbf{X})$ **then**
    return `FAIL`
end if
return $I_{\min}(Z|\mathbf{X})$

Theorem 2. *For any $\mathbf{X} \subseteq \mathbf{V} \setminus \mathbf{Y}$ and $Z \in \mathbf{V} \setminus \mathbf{Y}$ satisfying $\mathbf{X} \prec Z \prec \mathbf{Y}$, given a partial-cause structure $\widehat{\mathcal{C}}$ for $(\mathbf{X}, Z, \mathbf{Y})$, let R be the result of Algorithm 1 with inputs $\mathbf{X}$, Z, $P(\mathbf{V})$, and $\mathbb{C}_P(\widehat{\mathcal{C}})$. Then, the influence indicator $I(Z|\mathbf{X})$ is discernible from $\mathbb{M}_{\widehat{\mathcal{C}}}$ and $P(\mathbf{V})$ if and only if R is not* `FAIL`*. Moreover, if R is not* `FAIL`*, then $R = I^{\mathcal{M}^*}(Z|\mathbf{X})$.*

The proof is provided in Appendix A.5. When the common-cause structure is known exactly, applying Theorem 2 to the singleton set $\{\mathcal{C}\}$ recovers Proposition 2. Importantly, Theorem 2 does not require determining all common causes or a unique adjustment set. Different candidate structures may imply different adjustments yet yield the same influence indicator. An example of the discernibility with a partial-cause structure is given in Appendix C.2. When candidates disagree, additional structural knowledge can narrow the candidate set and help make the indicator discernible.

4.2 An active approach to influence discernibility

The preceding results characterize discernibility given structural information, but do not specify how it can be acquired efficiently. When candidates disagree and Algorithm 1 returns `FAIL`, the question is how to actively enrich the available knowledge until the influence indicator becomes discernible. To this end, we elicit additional knowledge from an expert through queries about common causes. Compared to requiring the expert to enumerate all relevant common causes in advance, it is easier to query a specified variable directly. As each query incurs a cost, we seek to acquire common-cause knowledge to achieve discernibility in a cost-effective manner.

Starting from a partial-cause structure $\widehat{\mathcal{C}}$ for $(\mathbf{X}, Z, \mathbf{Y})$, possibly empty, we ask whether a selected variable W belongs to the common-cause structure $\mathcal{C}$. The query thus concerns common causes of Z and $\mathbf{X}$ or $\mathbf{Y}$. An affirmative answer adds the variable W to $\widehat{\mathcal{C}}$, whereas a negative answer adds it to an exclusion set $\mathcal{E}$. The candidates consistent with the answers are $\mathbb{C}^{\mathcal{E}}(\widehat{\mathcal{C}}) := \{\mathcal{C} \in \mathbb{C}(\widehat{\mathcal{C}}) : \mathcal{C} \cap \mathcal{E} = \emptyset\}$. We design a heuristic to guide the next query using conditional mutual information. Specifically, we define the following two scores: $a_W := I(W; Z|\widehat{\mathcal{C}})$ and $b_W := I(W; \mathbf{X} \cup \mathbf{Y}|\widehat{\mathcal{C}})$, both computable from $P(\mathbf{V})$, where I denotes conditional mutual information. To place them on comparable scales, we normalize by conditional entropy H, obtaining $\tilde{a}_W := a_W / H(Z|\widehat{\mathcal{C}})$ and $\tilde{b}_W := b_W / H(\mathbf{X} \cup \mathbf{Y}|\widehat{\mathcal{C}})$. The final score $\omega(W)$ is defined as the harmonic mean of $\tilde{a}_W$ and $\tilde{b}_W$. Intuitively, this score favors variables informative about both the action and the outcomes beyond the observed contexts and known common causes, rather than those strongly associated with only one side.

Since the context $\mathbf{X}$ is already conditioned on, variables in $\mathbf{X}$ need not be queried for membership in $\mathcal{C}$. We thus select the variable W maximizing $\omega(W)$ from $\mathbf{V} \setminus (\mathbf{X} \cup Z \cup \mathbf{Y} \cup \widehat{\mathcal{C}} \cup \mathcal{E})$, excluding the actionable variable, the outcomes, and previously queried variables. As summarized in Algorithm 2, the procedure alternates between querying and discernibility checking until the indicator is determined or the query budget is exhausted. Given the observational distribution $P(\mathbf{V})$ and correct expert responses, the true common-cause structure remains among the candidates throughout the procedure. Hence, by Theorem 2, any returned indicator equals $I^{\mathcal{M}^*}(Z|\mathbf{X})$, while `FAIL` indicates disagreement among the candidates. With a sufficient budget, after at most $|\mathbf{V} \setminus (\mathbf{X} \cup \{Z\} \cup \mathbf{Y})|$ distinct queries, all common causes outside $\mathbf{X}$ are determined. The remaining candidates can differ only in whether variables already contained in $\mathbf{X}$ also belong to $\mathcal{C}$, which leaves the adjustment set $\mathcal{C} \setminus \mathbf{X}$ unchanged.

Algorithm 2 Discerning with active queries

Input: context $\mathbf{X}$; actionable variable Z; $\mathcal{L}_O$ data $P(\mathbf{V})$; query budget B;
Output: $I^{\mathcal{M}^*}(Z|\mathbf{X})$ if discerned, else `FAIL`.

$\widehat{\mathcal{C}} \leftarrow \emptyset$; $\mathcal{E} \leftarrow \emptyset$
$R \leftarrow$ run Algorithm 1 over $\mathbb{C}^{\mathcal{E}}(\widehat{\mathcal{C}})$
while $R =$ `FAIL` **and** $B > 0$ **do**
  $W^* \leftarrow \arg\max_W \omega(W)$
  Query whether W^* belongs to $\mathcal{C}$;
  $B \leftarrow B - 1$;
  if yes **then** $\widehat{\mathcal{C}} \leftarrow \widehat{\mathcal{C}} \cup \{W^*\}$; **else** $\mathcal{E} \leftarrow \mathcal{E} \cup \{W^*\}$
  $R \leftarrow$ run Algorithm 1 over $\mathbb{C}^{\mathcal{E}}(\widehat{\mathcal{C}})$
end while
return R

5 Experiments

In this section, we provide empirical evidence for our theoretical analysis by examining whether individual influence indicators can be correctly determined before the corresponding causal effects are accurately estimated. We also evaluate the effectiveness of the query selection strategies.

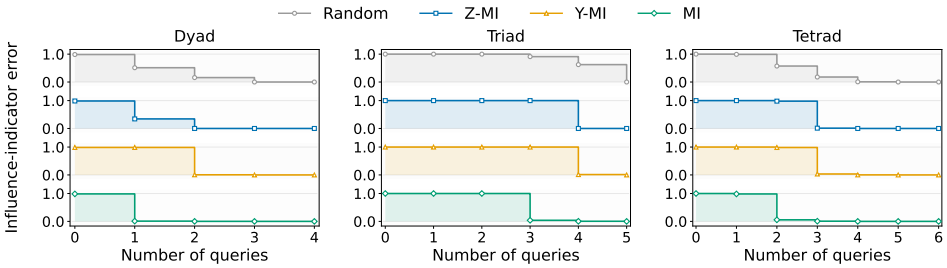

Figure 3: Influence-indicator mistake rates versus the number of queries on synthetic tasks.

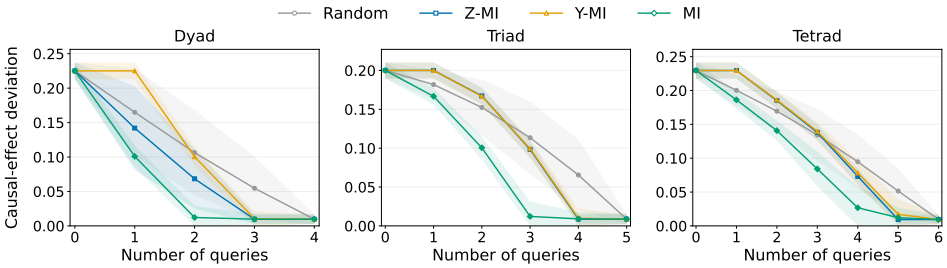

Figure 4: Causal-effect deviations versus the number of queries on synthetic tasks.

Table 1: Queries required for accurately assessing influence indicators (Q_R) and causal effects (Q_I).

Method	DYAD		TRIAD		TETRAD	
	Q_R	Q_I	Q_R	Q_I	Q_R	Q_I
RANDOM	1.66 ± 0.75	3.49 ± 0.85	4.54 ± 0.65	4.64 ± 0.73	2.75 ± 0.77	5.70 ± 0.71
Z-MI	1.33 ± 0.50	2.87 ± 0.80	4.00 ± 0.00	4.14 ± 0.51	2.99 ± 0.18	5.17 ± 0.55
Y-MI	1.98 ± 0.23	3.19 ± 0.57	4.01 ± 0.11	4.16 ± 0.52	3.02 ± 0.24	5.28 ± 0.61
MI	1.00 ± 0.15	2.29 ± 0.86	3.04 ± 0.22	3.25 ± 0.79	2.05 ± 0.30	4.52 ± 0.91

Tasks. We simulate three synthetic tasks (DYAD, TRIAD, and TETRAD) and a real-world case study (BERMUDA). We repeat each experiment ten times, with samples drawn from the underlying SCM. At each query step, we evaluate the candidates using observational data, and update the candidate set according to the expert response. More detailed settings are given in Appendix D.

Baselines. We compare four strategies for selecting the next variable to query: (1) RANDOM, which selects an unqueried variable uniformly at random; (2) Z-MI, which selects the variable with the highest conditional mutual information with the actionable variable; (3) Y-MI, which selects the variable with the highest conditional mutual information with the outcome variables; and (4) MI, which uses the harmonic mean of the two normalized scores described in Section 4.2. All scores are conditioned on the currently confirmed common causes and computed from observational data.

Fig. 3 shows the convergence of the influence indicators as common-cause knowledge is acquired. Each strategy is displayed in a separate lane, with shading between the error curve and zero. Across all three tasks, MI reduces the indicator mistake rate with fewer queries than the other baselines, while Z-MI and Y-MI tend to undesirably select variables associated with the action and the outcomes, without necessarily selecting relevant common causes. Fig. 4 shows that causal-effect deviations generally decrease as the number of queries increases. Notably, the deviations continue to diminish after the indicator mistake rates in Fig. 3 have become small. Thus, further queries can improve causal-effect estimation even when the influence judgments are already accurate. This empirically supports our theoretical insight: limited common-cause knowledge can provide useful decision guidance before the underlying causal effects are precisely estimated.

We further compare the number of queries required by the different strategies in Table 1, reporting means and standard deviations over the repetitions. Across all three tasks, MI achieves lower mean query counts for both Q_R and Q_I compared to the other baselines. Moreover, the mean Q_R is smaller than the mean Q_I for every strategy and task, with particularly pronounced gaps on DYAD and TETRAD. These results further support that correct influence judgments can be obtained with

fewer queries than are needed for precise causal-effect estimates. We note that the task names refer to the number of common causes, rather than the difficulty of attaining discernibility. Since an indicator can be determined without knowing all common causes, more common causes do not necessarily require more queries. Together, these results demonstrate the efficacy of the proposed query selection strategies in reducing the expert knowledge needed for rehearsal decisions.

We further conduct experiments in the BERMUDA task, with total alkalinity as the actionable variable in Fig. 5. The indicator mistake rate reaches zero after one query with MI and three queries with RANDOM. Notably, for both strategies, the causal-effect estimates have not fully converged when the indicator mistake rate reaches zero. Thus, the separation between correct influence judgments and precise causal-effect estimation is observed under both active and random querying. This is consistent with our previous analysis: limited common-cause knowledge can already provide useful decision guidance without precisely determining the corresponding causal effects. The full results across all actionable variables and query strategies are provided in Appendix E.

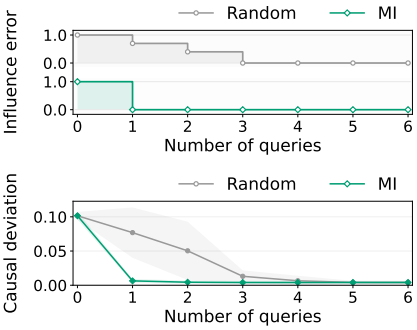

Figure 5: Influence-indicator errors and causal-effect deviations in BERMUDA.

6 RELATED WORK

Rehearsal learning. The rehearsal paradigm advocates anticipating the desired influence of actions on future outcomes in a hypothetical world (Zhou, 2022; 2023). Under this paradigm, a series of rehearsal learning techniques have been developed for the AUF problem. Qin et al. (2023) proposed the first rehearsal learning method that jointly optimizes decision actions and learns structural rehearsal models capable of characterizing reciprocal interactions from interactive data. Subsequent studies developed this direction for non-stationary environments (Du et al., 2024), nonlinear settings (Qin et al., 2025), sequential decision contexts (Tao et al., 2025), and long-term scenarios (Du et al., 2025b). Complementing these advances, Tao et al. (2026a) established a measure of influence quantifying the capacity of an actionable variable in avoiding undesired outcomes. More recently, it was shown that the order structure is more essential than the graph structure for AUF in rehearsal learning (Tao et al., 2026b). These studies demonstrate the empirical effectiveness of rehearsal learning in various AUF scenarios, while its theoretical feasibility remains under-explored.

Causality. Inferring causal models from data is a fundamental task in the literature. Depending on the available information, causal quantities are organized into observational, interventional, and counterfactual levels, forming the causal hierarchy (Shpitser & Pearl, 2008). A substantial body of work has been devoted to characterizing this hierarchy and identifying higher-level quantities from lower-level information (Zhang & Hyvärinen, 2009; Peters et al., 2014; Shimizu, 2014; Ibeling & Icard, 2020; Xia et al., 2021; Yang & Bareinboim, 2025; Li et al., 2026). Beyond point identification, another line of work studies partial identification of causality (Tian & Pearl, 2000; He & Geng, 2008; Zhang et al., 2022; Park & Lee, 2026; Choe et al., 2026). This work studies the influence judgments needed for rehearsal decisions and their discernibility under limited structural knowledge.

7 CONCLUSION

In this paper, we establish a formal framework for understanding the rehearsal decision in relation to statistical prediction and causal identification. We show that rehearsal requires information beyond the observational level but less than the interventional level. This establishes that solving the AUF problem does not, in general, require complete causal identification. We further characterize when the desired influence is discernible from observational data under structural knowledge, showing that common-cause knowledge suffices and that influence discernibility under partial common-cause knowledge can be determined with a principled procedure. Finally, we develop a novel strategy for acquiring partial knowledge, and empirically demonstrate that influence indicators can be determined with fewer queries before causal-effect estimates converge. Future work includes extending the influence discernibility to multiple actionable variables and accommodating latrent confounding.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, and clean and reformat datasets. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, or interpret results, and support for qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to edit the paper for readability. We have reviewed all AI-assisted work. We checked translations and reformatted datasets before use, and reviewed readability edits so that technical claims remained unchanged. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3155–3164, 2020.

Andreas Andersson and Nicholas Bates. In situ measurements used for coral and reef-scale calcification structural equation modeling including environmental and chemical measurements, and coral calcification rates in bermuda from 2010 to 2012 (beacon project), 2018.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty in artificial intelligence (UAI)*, pp. 46–54. Elsevier, 1994.

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

Christopher M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

David Maxwell Chickering, Dan Geiger, and David Heckerman. Learning bayesian networks: Search methods and experimental results. In *Pre-proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, volume R0, pp. 112–128. PMLR, 1995.

Yesong Choe, Yeahoon Kwon, Min Woo Park, and Sanghack Lee. Canonical domain reduction for partial counterfactual identification. In *Forty-Second Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2026.

Travis A Courtney, Mario Lebrato, Nicholas R Bates, Andrew Collins, Samantha J De Putron, Rebecca Garley, Rod Johnson, Juan-Carlos Molinero, Timothy J Noyes, Christopher L Sabine, et al. Environmental controls on modern scleractinian coral and reef-scale calcification. *Science advances*, 3(11):e1701356, 2017.

Wen-Bo Du, Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Avoiding undesired future with minimal cost in non-stationary environments. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 135741–135769, 2024.

Wen-Bo Du, Hao-Yi Lei, Lue Tao, Tian-Zuo Wang, and Zhi-Hua Zhou. Enabling optimal decisions in rehearsal learning under care condition. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pp. 14536–14561, 2025a.

Wen-Bo Du, Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Variance-reduced long-term rehearsal learning with quadratic programming reformulation. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 113245–113275, 2025b.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9:2523–2547, 2008.

Duligur Ibeling and Thomas Icard. Probabilistic reasoning across the causal hierarchy. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pp. 10170–10177, 2020.

Yiqing Li, Xiaofei Wang, Boyang Sun, Yewei Xia, and Kun Zhang. Independence test for linear non-gaussian data and applications in causal discovery. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Min Woo Park and Sanghack Lee. Counterfactual structural causal bandits. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. URL `http://jmlr.org/papers/v15/peters14a.html`.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Rehearsal learning for avoiding undesired future. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 80517–80542, 2023.

Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Gradient-based nonlinear rehearsal learning with multivariate alterations. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 26859–26867, 2025.

Arvind Raghavan and Elias Bareinboim. Counterfactual realizability. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Shohei Shimizu. LiNGAM: Non-Gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Ilya Shpitser and Judea Pearl. What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 352–359, 2007.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Lue Tao, Tian-Zuo Wang, Yuan Jiang, and Zhi-Hua Zhou. Avoiding undesired future with sequential decisions. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 6245–6253, 2025.

Lue Tao, Tian-Zuo Wang, Yuan Jiang, and Zhi-Hua Zhou. On measuring influence in avoiding undesired future. In *International Conference on Learning Representations (ICLR)*, 2026a.

Yu-Xuan Tao, Tian-Zuo Wang, and Zhi-Hua Zhou. Order-based rehearsal learning. *arXiv preprint arXiv:2605.04955*, 2026b.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1 edition, 1995.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Hongshuo Yang and Elias Bareinboim. A hierarchy of graphical models for counterfactual inferences. *Advances in Neural Information Processing Systems*, 38:130941–130988, 2025.

Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In *International Conference on Machine Learning (ICML)*, pp. 26548–26558. PMLR, 2022.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

Zhi-Hua Zhou. Rehearsal: Learning from prediction to decision. *Frontiers of Computer Science*, 16 (4):164352, 2022.

Zhi-Hua Zhou. Rehearsal: Learning from prediction to decision. Keynote at the CCF Conference on AI, Urumqi, China, 2023.

A PROOFS

A.1 PROOF OF THEOREM 1

Proof. We prove the two directions separately. Fix $\mathcal{M}^* = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ and write $\mathbf{A} := \mathbf{V} \setminus \mathbf{Y}$. For an action $\mathbf{z} \in \Delta_{\mathbf{Z}}$, let $\mathbf{Z}' := \{Z \in \mathbf{Z} : z_Z \neq \bot\}$ and let $\mathbf{z}'$ be the restriction of $\mathbf{z}$ to $\mathbf{Z}'$. Potential outcomes indexed by $\mathbf{z}$ are thus those under the alteration of $\mathbf{Z}'$ to $\mathbf{z}'$.

Positive direction ($\mathcal{L}_I$ implies $\mathcal{L}_R$). Suppose $\mathcal{M} \sim_I \mathcal{M}^*$. Consider any disjoint set of variables $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{A}$ and any action $\mathbf{z} \in \Delta_{\mathbf{Z}}$. According to Definitions 3 and 5, the two models $\mathcal{M}$ and $\mathcal{M}^*$ agree on the interventional distribution:

$$P^{\mathcal{M}}(\mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{Y}_{\mathbf{z}} = \mathbf{y}) = P^{\mathcal{M}^*}(\mathbf{X}_{\mathbf{z}} = \mathbf{x}, \mathbf{Y}_{\mathbf{z}} = \mathbf{y})$$

for every $(\mathbf{x}, \mathbf{y}) \in \Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}$. Marginalizing over $\mathbf{Y}_{\mathbf{z}}$ shows that the two models also agree on the distribution of $\mathbf{X}_{\mathbf{z}}$. Consequently, their influence indicators in Definition 4 have the same domain, and on this domain their rehearsal probabilities satisfy

$$P^{\mathcal{M}}(\mathbf{Y}_{\mathbf{z}} \in \mathcal{S} \mid \mathbf{X}_{\mathbf{z}} = \mathbf{x}) = P^{\mathcal{M}^*}(\mathbf{Y}_{\mathbf{z}} \in \mathcal{S} \mid \mathbf{X}_{\mathbf{z}} = \mathbf{x}).$$

Meanwhile, by Definition 4, the influence indicator is uniquely determined by the corresponding rehearsal probability, that is,

$$I^{\mathcal{M}}(\mathbf{Z} = \mathbf{z} \mid \mathbf{X} = \mathbf{x}) = \mathbb{1}\left[P^{\mathcal{M}}(\mathbf{Y}_{\mathbf{z}} \in \mathcal{S} \mid \mathbf{X}_{\mathbf{z}} = \mathbf{x}) \geq P^{\mathcal{M}}(\mathbf{Y}_{\mathbf{z}} \notin \mathcal{S} \mid \mathbf{X}_{\mathbf{z}} = \mathbf{x})\right].$$

Therefore, we have $\mathcal{M} \sim_R \mathcal{M}^*$.

Negative direction ($\mathcal{L}_R$ does not imply $\mathcal{L}_I$). We prove this by constructing a recursive SCM that preserves every influence indicator of $\mathcal{M}^*$ but changes at least one interventional probability.

First, fix a topological order of $\mathcal{M}^*$. For each $(\mathbf{a}, \mathbf{u}) \in \Omega_{\mathbf{A}} \times \Omega_{\mathbf{U}}$, we recursively define auxiliary functions h_V in this order by

$$h_V(\mathbf{a}, \mathbf{u}) := \begin{cases} a_V, & V \in \mathbf{A}, \\ f_V((h_W(\mathbf{a}, \mathbf{u}))_{W \in \mathbf{Pa}_V}, u_V), & V \in \mathbf{Y}. \end{cases}$$

For each $A \in \mathbf{A}$, we define

$$g_A(\mathbf{a}, \mathbf{u}) := f_A((h_W(\mathbf{a}, \mathbf{u}))_{W \in \mathbf{Pa}_A}, u_A).$$

These definitions substitute the original mechanisms of $\mathbf{Y}$ into the mechanism for A, while retaining variables in $\mathbf{A}$ as arguments. Consequently, g_A depends only on the original noises and on variables in $\mathbf{A}$ preceding A in the topological order.

We construct the model $\mathcal{M}_{\mathbf{s},\epsilon}$ from $\mathcal{M}^*$ and two well chosen terms $\mathbf{s} \in \Omega_{\mathbf{Y}}$ and $\epsilon \in (0, 1)$.

Let $B \sim \mathrm{Bern}(\epsilon)$ independently of $\mathbf{U}$. We define $\mathcal{M}_{\mathbf{s},\epsilon}$ by the structural equations

$$A \leftarrow g_A(\mathbf{a}, \mathbf{u}), \qquad A \in \mathbf{A},$$
$$Y \leftarrow \begin{cases} f_Y(\mathbf{pa}_Y, u_Y), & B = 0, \\ s_Y, & B = 1, \end{cases} \qquad Y \in \mathbf{Y}.$$

The noise variables of $\mathcal{M}_{\mathbf{s},\epsilon}$ are $\mathbf{U} \cup B$, with joint distribution $P(\mathbf{U})P(B)$. Every structural function depends only on endogenous variables preceding it in the topological order. Thus, $\mathcal{M}_{\mathbf{s},\epsilon}$ is recursive.

For any action $\mathbf{z}$ on $\mathbf{A}$, induction along the topological order shows that $\mathbf{A}_{\mathbf{z}}$ agrees with that in $\mathcal{M}^*$ in both branches, while the branch $B = 0$ reproduces the original solution. Hence, for every $(\mathbf{u}, b)$,

$$\mathbf{A}_{\mathbf{z}}^{\mathcal{M}_{\mathbf{s},\epsilon}}(\mathbf{u}, b) = \mathbf{A}_{\mathbf{z}}^{\mathcal{M}^*}(\mathbf{u}),$$
$$\mathbf{Y}_{\mathbf{z}}^{\mathcal{M}_{\mathbf{s},\epsilon}}(\mathbf{u}, b) = \begin{cases} \mathbf{Y}_{\mathbf{z}}^{\mathcal{M}^*}(\mathbf{u}), & b = 0, \\ \mathbf{s}, & b = 1. \end{cases}$$

In particular, for any disjoint $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{A}$, any action $\mathbf{z} \in \Delta_{\mathbf{Z}}$, and every $\mathbf{x}$,

$$P^{\mathcal{M}_{\mathbf{s},\epsilon}}(\mathbf{X}_{\mathbf{z}} = \mathbf{x}) = P^{\mathcal{M}^*}(\mathbf{X}_{\mathbf{z}} = \mathbf{x}).$$

Thus, the two rehearsal valuations have the same query domain.

Let $\mathcal{Q}$ be this domain, consisting of the queries $q = (\mathbf{X}, \mathbf{Z}, \mathbf{x}, \mathbf{z})$ with $P^{\mathcal{M}^*}(\mathbf{X_z} = \mathbf{x}) > 0$. The set $\mathcal{Q}$ is finite because $\mathbf{V}$ and all variable domains are finite. For each $q \in \mathcal{Q}$, we write

$$p_q := P^{\mathcal{M}^*}(\mathbf{Y_z} \in \mathcal{S} \mid \mathbf{X_z} = \mathbf{x}),$$
$$p_q^\epsilon := P^{\mathcal{M}_{\mathbf{s},\epsilon}}(\mathbf{Y_z} \in \mathcal{S} \mid \mathbf{X_z} = \mathbf{x}).$$

Since $\mathbf{X_z}$ depends only on $\mathbf{U}$, it is independent of B. Therefore, we have

$$p_q^\epsilon = (1-\epsilon)p_q + \epsilon\, \mathbb{1}[\mathbf{s} \in \mathcal{S}].$$

Two cases remain. *Case 1:* $p_q = 1$, for every $q \in \mathcal{Q}$. In this case, we choose $\mathbf{s} \notin \mathcal{S}$ and $\epsilon = 1/4$. Then, we can derive that $p_q^\epsilon = 3/4$, for every $q \in \mathcal{Q}$, Thus, we have $p_q \neq p_q^\epsilon$ and $\mathcal{M}_{\mathbf{s},\epsilon} \not\sim_I \mathcal{M}^*$. Meanwhile, both probabilities p_q and p_q^ϵ are larger than $1/2$. By Definition 4, we have $I^{\mathcal{M}^*}(\mathbf{Z} = \mathbf{z}|\mathbf{X} = \mathbf{x}) = I^{\mathcal{M}_{\mathbf{s},\epsilon}}(\mathbf{Z} = \mathbf{z}|\mathbf{X} = \mathbf{x})$ for every $q \in \mathcal{Q}$. Thus, we have $\mathcal{M}_{\mathbf{s},\epsilon} \sim_R \mathcal{M}^*$.

Case 2: $p_q < 1$, for some $q \in \mathcal{Q}$. In this case, we choose $\mathbf{s} \in \mathcal{S}$. Then, we can derive that $p_q^\epsilon = p_q + \epsilon(1-p_q)$. If $p_q \geq 1 - p_q$, we have $p_q^\epsilon \geq p_q \geq 1 - p_q \geq 1 - p_q^\epsilon$, which implies that the influence indicator is the same. If $p_q < 1 - p_q$, we have that the continuity in ϵ yields $p_q^\epsilon < 1 - p_q^\epsilon$ for a sufficiently small $\epsilon > 0$. In either subcase, p_q and p_q^ϵ entail the same influence indicator, i.e., $I^{\mathcal{M}^*}(\mathbf{Z} = \mathbf{z}|\mathbf{X} = \mathbf{x}) = I^{\mathcal{M}_{\mathbf{s},\epsilon}}(\mathbf{Z} = \mathbf{z}|\mathbf{X} = \mathbf{x})$. Moreover, as there exists some query $q \in \mathcal{Q}$ with $p_q < 1$, the rehearsal probability $p_q^\epsilon = p_q + \epsilon(1-p_q)$ strictly increases, i.e., $p_q^\epsilon > p_q$. Therefore, we obtain $\mathcal{M}_{\mathbf{s},\epsilon} \sim_R \mathcal{M}^*$, while $p_q^\epsilon \neq p_q$ for at least one query q, which means that $\mathcal{M}_{\mathbf{s},\epsilon} \not\sim_I \mathcal{M}^*$.

As $\mathcal{M}^*$ is arbitrary, the negative direction holds relative to every $\mathcal{M}^* \in \mathbb{M}$. □

A.2 Proof of Corollary 1

The following definition gives the third level of the causal hierarchy (Pearl & Mackenzie, 2018).

Definition 10 (Counterfactual Valuation)**.** An SCM $\mathcal{M}$ induces $\mathcal{L}_C(\mathcal{M})$, a collection of probability mass functions over counterfactual events $(\mathbf{Z}_1)_{\mathbf{x}_1} = \mathbf{z}_1, \ldots, (\mathbf{Z}_k)_{\mathbf{x}_k} = \mathbf{z}_k$ for all pairs of sets $\mathbf{X}_i, \mathbf{Z}_i \subseteq \mathbf{V}$ with $\mathbf{X}_i \cap \mathbf{Z}_i = \emptyset$, $\forall i \in \{1, \ldots, k\}$. For each assignment $\mathbf{x}_i \in \Omega_{\mathbf{X}_i}$ and $\mathbf{z}_i \in \Omega_{\mathbf{Z}_i}$, $i \in \{1, \ldots, k\}$, the probability is given by

$$P^{\mathcal{M}}\big((\mathbf{Z}_1)_{\mathbf{x}_1} = \mathbf{z}_1, \ldots, (\mathbf{Z}_k)_{\mathbf{x}_k} = \mathbf{z}_k\big) := \sum_{\{\mathbf{u} \,:\, (\mathbf{Z}_i)_{\mathbf{x}_i}(\mathbf{u}) = \mathbf{z}_i \;\forall i\}} P(\mathbf{u}).$$

The interventional valuation of Definition 3 is the special case $k = 1$. Following Bareinboim et al. (2022), the layers L_1, L_2, L_3 in the causal hierarchy theorem (CHT) correspond to our valuations $\mathcal{L}_O, \mathcal{L}_I, \mathcal{L}_C$, respectively, and their notion of collapse corresponds to the implication in Definition 6.

Fact 1 (CHT, Bareinboim et al., 2022, Thm. 27.1)**.** *For each $(i, j) \in \{(O, I), (O, C), (I, C)\}$, with respect to the Lebesgue measure over (a suitable encoding of L_3-equivalence classes of) SCMs, the set of $\mathcal{M}^* \in \mathbb{M}$ relative to which $\mathcal{L}_i$ implies $\mathcal{L}_j$ has measure zero.*

Proof. There are six ordered pairs to check. The three not involving R, namely (O, I), (O, C), and (I, C), follow directly from Fact 1.

(R, I). By the negative direction of Theorem 1, $\mathcal{L}_R$ does not imply $\mathcal{L}_I$ relative to any $\mathcal{M}^*$.

(R, C). We show this by contradiction. Suppose that $\mathcal{L}_R$ implies $\mathcal{L}_C$ relative to some $\mathcal{M}^*$. Since $\mathcal{L}_C$ determines $\mathcal{L}_I$ by taking $k = 1$ in Definition 10, transitivity implies that $\mathcal{L}_R$ implies $\mathcal{L}_I$ relative to the same $\mathcal{M}^*$, contradicting Theorem 1.

(O, R). Fix any $\mathcal{M}^*$, and write $\mathbf{A} := \mathbf{V} \setminus \mathbf{Y}$. Since $|\Omega_{\mathbf{A}}| \geq 3$, there exists $\mathbf{a} \in \Omega_{\mathbf{A}}$ with $P^{\mathcal{M}^*}(\mathbf{A} = \mathbf{a}) \leq 1/3$. Also, we choose a desired outcome $\mathbf{y}_+ \in \mathcal{S}$ and an undesired outcome $\mathbf{y}_- \in \Omega_{\mathbf{Y}} \setminus \mathcal{S}$.

We next construct two recursive SCMs $\mathcal{M}_+$ and $\mathcal{M}_-$ that are both $\mathcal{L}_O$-consistent but are not both $\mathcal{L}_R$-consistent with $\mathcal{M}^*$. Let exogenous noise $\mathbf{U} = (\mathbf{U_A}, \mathbf{U_Y})$ with joint distribution

$$P(\mathbf{U_A} = \mathbf{a}, \mathbf{U_Y} = \mathbf{y}) := P^{\mathcal{M}^*}(\mathbf{A} = \mathbf{a}, \mathbf{Y} = \mathbf{y}).$$

We define two recursive SCMs $\mathcal{M}_+$ and $\mathcal{M}_-$ respectively through the following equations

$$\mathbf{A} \leftarrow \mathbf{U_A}, \qquad \mathbf{Y} \leftarrow \begin{cases} \mathbf{U_Y}, & \mathbf{A} = \mathbf{U_A}, \\ \mathbf{y}_\pm, & \mathbf{A} \neq \mathbf{U_A}. \end{cases}$$

Without alteration, both models yield $(\mathbf{A}, \mathbf{Y}) = (\mathbf{U_A}, \mathbf{U_Y})$. Hence, we have $\mathcal{M}_+ \sim_O \mathcal{M}^*$ and $\mathcal{M}_- \sim_O \mathcal{M}^*$. Under the alteration of setting $\mathbf{A}$ to $\mathbf{a}$, we have

$$P^{\mathcal{M}_-}(\mathbf{Y_a} \in \mathcal{S}) = P^{\mathcal{M}^*}(\mathbf{A} = \mathbf{a}, \mathbf{Y} \in \mathcal{S}) \leq \frac{1}{3},$$
$$P^{\mathcal{M}_+}(\mathbf{Y_a} \in \mathcal{S}) = 1 - P^{\mathcal{M}^*}(\mathbf{A} = \mathbf{a}, \mathbf{Y} \notin \mathcal{S}) \geq \frac{2}{3}.$$

The rehearsal query with $\mathbf{X} = \emptyset$ and $\mathbf{Z} = \mathbf{A}$ is well-defined under Definition 4 in every model. Consequently, we obtain $I^{\mathcal{M}_-}(\mathbf{A} = \mathbf{a}) = 0$, and $I^{\mathcal{M}_+}(\mathbf{A} = \mathbf{a}) = 1$.

Therefore, at least one of $\mathcal{M}_-$ and $\mathcal{M}_+$ is $\mathcal{L}_O$-consistent but not $\mathcal{L}_R$-consistent with $\mathcal{M}^*$.

Since $\mathcal{M}^*$ is arbitrary, $\mathcal{L}_O$ does not imply $\mathcal{L}_R$ relative to any $\mathcal{M}^*$.

Together with Fact 1, these three cases establish the corollary. □

A.3 Proof of Proposition 1

Proof. Fix a context $\mathbf{x}$ and an action z. For $z = \perp$, the rehearsal probability $P^{\mathcal{M}}(\mathbf{Y}_z \in \mathcal{S} | \mathbf{X} = \mathbf{x})$ is the conditional probability $P(\mathbf{Y} \in \mathcal{S} | \mathbf{X} = \mathbf{x})$. For $z \in \Omega_Z$, consider any $\mathcal{M} \in \mathbb{M}_\mathcal{G}$ inducing $P(\mathbf{V})$. Since $\mathbf{X} \prec Z$, altering Z leaves $\mathbf{X}$ unchanged. The Markovian factorization therefore gives

$$P^{\mathcal{M}}(\mathbf{Y}_z \in \mathcal{S} \mid \mathbf{X} = \mathbf{x}) = \frac{1}{P(\mathbf{X} = \mathbf{x})} \sum_{\substack{\mathbf{v} \in \Omega_\mathbf{V}: \\ v_Z = z,\ \mathbf{v_X} = \mathbf{x},\ \mathbf{v_Y} \in \mathcal{S}}} \prod_{V \in \mathbf{V} \setminus \{Z\}} P(v_V \mid \mathbf{v}_{\mathbf{Pa}_V}). \tag{1}$$

Positivity ensures that every factor on the right is well-defined. Since the right-hand side depends only on $\mathcal{G}$ and $P(\mathbf{V})$, every model in $\mathbb{M}_\mathcal{G}$ that induces $P(\mathbf{V})$ has the same rehearsal probability. By Definition 4, this yields the same influence indicator for every model in $\mathbb{M}_\mathcal{G}$. □

A.4 Proof of Proposition 2

Proof. Consider any $\mathcal{M} \in \mathbb{M}_\mathcal{C}$ that induces $P(\mathbf{V})$. Let $\mathcal{G}$ denote the graph of $\mathcal{M}$. Write $\mathbf{D} := \mathbf{X} \cup \mathcal{C}$ for brevity. We next show that $\mathbf{D}$ blocks every path from Z to $\mathbf{Y}$ in $\mathcal{G}_{\underline{Z}}$.

Since $\mathbf{X} \prec Z$ and every variable in $\mathcal{C}$ is an ancestor of Z, no variable in $\mathbf{D}$ is a descendant of Z. Suppose a path from Z to $\mathbf{Y}$ in $\mathcal{G}_{\underline{Z}}$ is not blocked given $\mathbf{D}$. It starts with an edge into Z. If it has no collider, its fork is a common ancestor of Z and $\mathbf{Y}$, which belongs to $\mathbf{D}$. Therefore, the path is blocked, a contradiction. Otherwise, let K be its first collider when traversing from Z. The preceding collider-free segment has a fork Q with directed paths to both Z and K. For the path from Z to $\mathbf{Y}$ in $\mathcal{G}_{\underline{Z}}$ to be not blocked, K must belong to $\mathbf{D}$ or have a descendant in $\mathbf{D}$. If K or one of its descendants belongs to $\mathbf{X}$, then Q has a directed path to $\mathbf{X}$. Otherwise, K or one of its descendants belongs to $\mathcal{C}$. Since every variable in $\mathcal{C}$ has a directed path in $\mathcal{G}_{\underline{Z}}$ to $\mathbf{X}$ or $\mathbf{Y}$, Q does as well. In either case, the fork Q is an ancestor of Z and has a directed path to $\mathbf{X}$ or $\mathbf{Y}$. It therefore belongs to $\mathbf{X}$ or $\mathcal{C}$, and is a conditioned noncollider on the path, a contradiction.

Thus, $\mathbf{D}$ blocks every path from Z to $\mathbf{Y}$ in $\mathcal{G}_{\underline{Z}}$, which means that $\mathbf{D}$ satisfies the back-door criterion. Hence, we can apply the conditional back-door adjustment to derive $P^{\mathcal{M}}(\mathbf{Y}_z \in \mathcal{S} | \mathbf{X} = \mathbf{x})$ from observational distribution $P(\mathbf{V})$. This yields the same influence indicator for every model in $\mathbb{M}_\mathcal{C}$. □

A.5 Proof of Theorem 2

Proof. By Proposition 2, every $\mathcal{C} \in \mathbb{C}_P(\widehat{\mathcal{C}})$ determines an indicator $I^\mathcal{C}(Z|\mathbf{X})$ shared by all models in $\mathbb{M}_\mathcal{C}$ that induce $P(\mathbf{V})$. Suppose Algorithm 1 returns a value R. Then, every candidate $\mathcal{C} \in \mathbb{C}_P(\widehat{\mathcal{C}})$ has the same indicator R. Consider any $\mathcal{M} \in \mathbb{M}_{\widehat{\mathcal{C}}}$ that induces $P(\mathbf{V})$, and let $\mathcal{C}_\mathcal{M}$ be its common-cause structure. Since $\mathcal{M}$ is consistent with $\widehat{\mathcal{C}}$ and induces $P(\mathbf{V})$, we have $\mathcal{C}_\mathcal{M} \in \mathbb{C}_P(\widehat{\mathcal{C}})$. By

Proposition 2, $I^{\mathcal{M}}(Z|\mathbf{X}) = I^{\mathcal{C}_{\mathcal{M}}}(Z|\mathbf{X}) = R$. Thus, all models in $\mathbb{M}_{\widehat{\mathcal{C}}}$ that induce $P(\mathbf{V})$ share the same indicator. In particular, since $\mathcal{M}^* \in \mathbb{M}_{\widehat{\mathcal{C}}}$, $R = I^{\mathcal{M}^*}(Z|\mathbf{X})$.

On the other hand, suppose Algorithm 1 returns FAIL. Then, there exist two candidates $\mathcal{C}_1, \mathcal{C}_2 \in \mathbb{C}_P(\widehat{\mathcal{C}})$ whose indicators differ at some observation $\mathbf{x}$ and action z. By definition of $\mathbb{C}_P(\widehat{\mathcal{C}})$, there exist models $\mathcal{M}_1 \in \mathbb{M}_{\mathcal{C}_1}$ and $\mathcal{M}_2 \in \mathbb{M}_{\mathcal{C}_2}$ that both induce $P(\mathbf{V})$. By Proposition 2, their indicators are $I^{\mathcal{C}_1}$ and $I^{\mathcal{C}_2}$, respectively, and they differ at $(\mathbf{x}, z)$. As both $\mathcal{M}_1$ and $\mathcal{M}_2$ belong to $\mathbb{M}_{\widehat{\mathcal{C}}}$, the disagreement implies that the influence indicator is not discernible. □

A.6 Incomparability of $\mathcal{L}_O$ and $\mathcal{L}_R$

The distinction between observational and rehearsal valuations indicates the gap between predicting outcomes and judging the influence of actions. The observational information is able to issue AUF warnings, while the rehearsal information supports decisions about how to resolve them. The order $O < R < I < C$ in Corollary 1 summarizes the corresponding directions of non-implication. The following proposition shows that $\mathcal{L}_R$ need not determine $\mathcal{L}_O$, reflecting its focus on influence judgments rather than exact observational probabilities.

Proposition 3. *There is an SCM $\mathcal{M}^*$ relative to which $\mathcal{L}_R$ does not imply $\mathcal{L}_O$.*

Proof. We construct two SCMs that agree on every influence indicator but differ in their observational distributions. Let $\mathbf{V} = \{Z, Y\}$, with both variables binary, $\mathbf{Y} = \{Y\}$, and $\mathcal{S} = \{1\}$. Consider the family of recursive SCMs with the following equations:

$$Z \leftarrow U_Z, \qquad Y \leftarrow ZU_1 + (1 - Z)U_0,$$

where U_Z, U_0, U_1 are mutually independent, $U_Z \sim \mathrm{Bern}(1/2)$, and $U_i \sim \mathrm{Bern}(p_i)$ for $i \in \{0, 1\}$. For every model in this family,

$$P(Y_z = 1) = P(Y = 1 \mid Z = z) = p_z, \qquad P(Y_\perp = 1) = (p_0 + p_1)/2.$$

Let $\mathcal{M}^*$ take $(p_0, p_1) = (0.2, 0.9)$ and $\mathcal{M}$ take $(p_0, p_1) = (0.3, 0.8)$.

In both models, $p_0 < 1/2 < p_1$ and $(p_0 + p_1)/2 = 0.55$. Thus, with empty context, the indicators for actions 0, 1, and $\perp$ are 0, 1, and 1, respectively, in both models. The query with both context and actionable sets empty also has indicator 1. With context $Z = z$, the actionable set must be empty, and the indicator is $\mathbb{1}[p_z \geq 1 - p_z]$, which again agrees between the two models. Consequently, we have $\mathcal{M} \sim_R \mathcal{M}^*$. In contrast, the observational distributions differ:

$$P^{\mathcal{M}}(Y = 1 \mid Z = 1) = 0.8 \neq 0.9 = P^{\mathcal{M}^*}(Y = 1 \mid Z = 1).$$

Therefore, we have $\mathcal{M} \not\sim_O \mathcal{M}^*$. This shows that $\mathcal{L}_R$ does not imply $\mathcal{L}_O$ relative to $\mathcal{M}^*$. □

B Adjustment Formula

In this section, we give the back-door adjustment that computes the rehearsal probability, and hence the influence indicator $I^{\mathcal{C}}(Z|\mathbf{X})$, from a common-cause structure $\mathcal{C}$ and $P(\mathbf{V})$.

Fix a common-cause structure $\mathcal{C}$ w.r.t. $(\mathbf{X}, Z, \mathbf{Y})$, and consider any $\mathcal{M} \in \mathbb{M}_{\mathcal{C}}$ inducing $P(\mathbf{V})$ with $P(\mathbf{v}) > 0$ for every $\mathbf{v} \in \Omega_{\mathbf{V}}$. Let $\mathbf{W} := \mathcal{C} \backslash \mathbf{X}$. As shown in Appendix A.4, $\mathbf{X} \cup \mathbf{W} = \mathbf{X} \cup \mathcal{C}$ contains no descendant of Z and blocks every back-door path from Z to $\mathbf{Y}$. Consequently, conditional on $\mathbf{X}$ and $\mathbf{W}$, the potential outcome $\mathbf{Y}_z$ is independent of Z. This yields

$$P^{\mathcal{M}}(\mathbf{Y}_z \in \mathcal{S} \mid \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}) = P(\mathbf{Y} \in \mathcal{S} \mid Z = z, \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}).$$

Averaging over $\mathbf{W}$ given $\mathbf{X} = \mathbf{x}$ gives the following adjustment formula:

$$p^{\mathcal{C}}_{z,\mathbf{x}} := \sum_{\mathbf{w} \in \Omega_{\mathbf{W}}} P(\mathbf{Y} \in \mathcal{S} \mid Z = z, \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}) \cdot P(\mathbf{W} = \mathbf{w} \mid \mathbf{X} = \mathbf{x}). \tag{2}$$

The right-hand side is determined by $P(\mathbf{V})$ and equals $P^{\mathcal{M}}(\mathbf{Y}_z \in \mathcal{S} | \mathbf{X} = \mathbf{x})$ for every model in $\mathbb{M}_{\mathcal{C}}$ that induces $P(\mathbf{V})$. When $\mathcal{C} \subseteq \mathbf{X}$, the sum reduces to $P(\mathbf{Y} \in \mathcal{S} | Z = z, \mathbf{X} = \mathbf{x})$. For the null action, no mechanism is altered, so we set $p^{\mathcal{C}}_{\perp,\mathbf{x}} := P(\mathbf{Y} \in \mathcal{S} | \mathbf{X} = \mathbf{x})$.

Since the desired and undesired events are complementary, the influence indicator is given by

$$I^{\mathcal{C}}(Z = z \mid \mathbf{X} = \mathbf{x}) = \mathbb{1}\left[p^{\mathcal{C}}_{z,\mathbf{x}} \geq 1 - p^{\mathcal{C}}_{z,\mathbf{x}}\right]. \tag{3}$$

This value equals $I^{\mathcal{M}}(Z = z | \mathbf{X} = \mathbf{x})$ for every $\mathcal{M} \in \mathbb{M}_{\mathcal{C}}$ inducing $P(\mathbf{V})$.

C Illustrative Examples

C.1 An example of the hierarchy between rehearsal and intervention

We supply Theorem 1 with two SCMs that agree at the observational and rehearsal levels, while disagreeing at the interventional level.

Example 1. Let $\mathbf{V} = \{Z, Y\}$ be binary variables, where Z indicates whether a patient takes a treatment and Y indicates whether the patient recovers; recovery is desired, i.e., $\mathcal{S} = \{1\}$. Consider the true SCM $\mathcal{M}^*$ and a rehearsal SCM $\mathcal{M}$, each with independent noise variables U_Z, U_Y satisfying $P(U_Z{=}1) = 1/2$ and $P(U_Y{=}1) = 3/4$, and differing only in the mechanism for Y:

$$\mathcal{F}^* = \begin{cases} Z \leftarrow U_Z \\ Y \leftarrow Z \cdot U_Y \end{cases} \qquad \mathcal{F} = \begin{cases} Z \leftarrow U_Z \\ Y \leftarrow Z \cdot \max(1 - U_Z,\, U_Y) \end{cases} .$$

The two SCMs disagree when $Z = 1$ yet $U_Z = 0$, an event that never occurs without intervention. Consequently, $\mathcal{M} \sim_O \mathcal{M}^*$. In particular, $P^{\mathcal{M}^*}(Y = 1) = P^{\mathcal{M}}(Y = 1) = 3/8$, so the warning $P(Y \notin \mathcal{S}) > P(Y \in \mathcal{S})$ is issued. The two SCMs do not, however, agree at the interventional level. Indeed, $P^{\mathcal{M}^*}(Y{=}1|do(Z{=}1)) = 3/4$, $P^{\mathcal{M}}(Y{=}1|do(Z{=}1)) = 7/8$, and hence $\mathcal{M} \not\sim_I \mathcal{M}^*$. Nevertheless, the two SCMs agree on the rehearsal level: $\mathcal{M} \sim_R \mathcal{M}^*$. Thus, whether rehearsal is conducted using $\mathcal{M}$ or $\mathcal{M}^*$, the agent reaches the same decision for the AUF problem. Specifically, since $\mathbf{V} \setminus \mathbf{Y} = \{Z\}$, the pair with nonempty actionable set admitted by Definition 4 is $\mathbf{X} = \emptyset$ and $\mathbf{Z} = \{Z\}$. Thus, $\mathcal{L}_R$ includes the three indicators, one for each action $\Delta_Z = \{0, 1, \perp\}$. Under both SCMs, inaction and withholding treatment leave recovery less likely than non-recovery, whereas administering the treatment shifts the odds in favor of recovery. The corresponding influence indicators in rehearsal therefore coincide: $I^{\mathcal{M}^*}(Z{=}\perp) = I^{\mathcal{M}}(Z{=}\perp) = 0$, $I^{\mathcal{M}^*}(Z{=}0) = I^{\mathcal{M}}(Z{=}0) = 0$, $I^{\mathcal{M}^*}(Z{=}1) = I^{\mathcal{M}}(Z{=}1) = 1$. Both models thus resolve the warning with the same prescription: administer the treatment.

C.2 An example of discernibility from partial-cause structure

We supply Section 4.1 with an example in which discernibility is certified beyond the reach of Proposition 2: the influence indicator is settled even though the causal effect is not.

Example 2. Let $\mathbf{V} = \{A, B, Z, Y\}$ be binary variables, where Y is the outcome with desired set $\mathcal{S} = \{1\}$ and Z is actionable with $\Delta_Z = \{0, 1, \perp\}$; take $\mathbf{X} = \emptyset$. Consider the true SCM $\mathcal{M}^*$ with mutually independent noises satisfying $P(U_A{=}1) = P(U_B{=}1) = 1/2$, $P(U_Z{=}1) = 4/5$, and $P(U_Y{=}1) = 1/4$, and the following structural equations ($\oplus$ denotes exclusive-or):

$$\mathcal{F}^* = \begin{cases} A \leftarrow U_A \\ B \leftarrow U_B \\ Z \leftarrow (A \vee B) \oplus U_Z \\ Y \leftarrow \big(Z \wedge (A \vee B)\big) \oplus U_Y \end{cases} .$$

The induced observational distribution is strictly positive, with $P(Y{=}1) = 13/40 < 1/2$, so the AUF warning is issued. The true common-cause structure is $\mathcal{C} = \{A, B\}$. Suppose initially no common cause is known, i.e., $\widehat{\mathcal{C}} = \emptyset$. The observational distribution also admits a Markovian SCM $\mathcal{M}_0$ with factorization $P(A, B, Z, Y) = P(Z)P(A|Z)P(B|A, Z)P(Y|A, B, Z)$, in which Z has no parents and hence the common-cause structure is empty. This model gives $P^{\mathcal{M}_0}(Y{=}1|do(Z{=}1)) = 13/28 < 1/2$, whereas $P^{\mathcal{M}^*}(Y{=}1|do(Z{=}1)) = 5/8 > 1/2$. Thus, the two observationally equivalent models disagree on whether setting $Z = 1$ resolves the warning, and $I(Z)$ is not discernible without common-cause knowledge.

Now suppose the agent learns that A is a common cause, while remaining unsure whether B is also one. With $\widehat{\mathcal{C}} = \{A\}$, the remaining candidates are $\{A\}$ and $\{A, B\}$. Both are compatible with the observational distribution: the latter is induced by $\mathcal{M}^*$, whereas the factorization $P(A, B, Z, Y) = P(A)P(Z|A)P(B|A, Z)P(Y|A, B, Z)$ can be realized by a Markovian SCM $\mathcal{M}_A$ in which B is a descendant of Z and the common-cause structure is $\{A\}$. Consequently, $\mathbb{C}_P(\widehat{\mathcal{C}}) = \{\{A\}, \{A, B\}\}$.

The two candidates still disagree on the causal effect of setting $Z = 1$:

$$P^{\mathcal{M}_A}(Y{=}1|do(Z{=}1)) = 11/20, \qquad P^{\mathcal{M}^*}(Y{=}1|do(Z{=}1)) = 5/8.$$

Nevertheless, both probabilities exceed $1/2$. Setting $Z = 0$ yields probability $1/4$ under either candidate, while inaction leaves it at $13/40$. The corresponding influence indicators therefore coincide: $I^{\{A\}}(Z{=}\perp) = I^{\{A,B\}}(Z{=}\perp) = 0$, $I^{\{A\}}(Z{=}0) = I^{\{A,B\}}(Z{=}0) = 0$, and $I^{\{A\}}(Z{=}1) = I^{\{A,B\}}(Z{=}1) = 1$. Hence, Algorithm 1 returns their common indicator and, by Theorem 2, $I(Z)$ is discernible from $\mathbb{M}_{\widehat{\mathcal{C}}}$ and $P(\mathbf{V})$. Knowledge of a single common cause thus enables the agent to resolve the warning by setting $Z = 1$, without determining whether B is also a common cause or identifying the causal effect of that action.

D Detailed Settings

Metric. After t queries, we estimate the influence indicator and causal effect using the currently confirmed common causes. We report the indicator mistake rate $e_R(t)$ and the absolute causal-effect deviation $e_I(t)$. An indicator error is recorded if the estimate is incorrect for any action. For binary actions, effect deviation is the absolute error in the causal effect; for multiple alteration states, it is the largest absolute error among the pairwise probability contrasts. We denote by Q_R the first query count at which Algorithm 1 returns the correct influence indicator for all actions, and by Q_I the first query count at which the effect deviation meets an appropriate tolerance. The initial state corresponds to $t = 0$. If the tolerance is not met within the query budget B, the corresponding count is recorded as $B + 1$. Queries continue through the full budget even after the tolerance is met.

D.1 Synthetic Tasks

The three synthetic tasks, Dyad, Triad, and Tetrad, contain two, three, and four common causes, respectively. They share the same structural form but differ in the number of common causes and the functional machenisms of the actionable variable Z and the outcome variable Y. Let $C_1, \ldots, C_m, A, B$ be mutually independent variables from $\mathrm{Bern}(0.5)$. Let U_Z and U_Y be uniform on $\{1, \ldots, 20\}$ and independent of each other and these variables. The underlying SCM is given by

$$G_m \coloneqq \mathbb{1}\left[\sum_{i=1}^{m} C_i \geq h_m\right], Z \leftarrow \mathbb{1}[U_Z \leq 2 + 16A(1 - G_m)], Y \leftarrow \mathbb{1}[U_Y \leq 4 + 12ZG_m + 2B].$$

The pairs of common causes and threshold gates (m, h_m) are $(2, 1)$, $(3, 2)$, and $(4, 2)$ for Dyad, Triad, and Tetrad, respectively. The variables $C_1, \ldots, C_m$ are common causes of Z and Y, whereas A affects Z and B affects Y. The queryable variables are $\{C_1, \ldots, C_m, A, B\}$, giving query budgets of 4, 5, and 6. The actionable variable Z is binary, with feasible alterations $\{0, 1\}$ and $\perp$ denoting inaction. The desired set is $\mathcal{S} = \{1\}$, and no individual context is given, i.e., $\mathbf{X} = \emptyset$.

Let g_m denote $P(G_m = 1)$. The natural desired-outcome probability is $0.25 + 0.06g_m$, while the probabilities under actions of altering Z to 0 and 1 are 0.25 and $0.25 + 0.60g_m$, respectively. Thus, the causal effects $P(Y_1 = 1) - P(Y_0 = 1)$ are 0.45, 0.30, and 0.4125. At the odds threshold one, the influence indicators are $I(Z = \perp) = I(Z = 0) = 0, I(Z = 1) = 1$. The effect-estimation tolerance is set to 0.02 for all three tasks, which is appropriate for the three tasks. For each task, we use 10,000 observational samples per repetition and repeat the experiment 1,000 times.

D.2 Bermuda

The Bermuda task is derived from a real-world scenario involving the management of net coral ecosystem calcification in Bermuda (Aglietti et al., 2020). The variables describe bottom light levels ($Light$), bottom temperature (Tem), sea surface salinity (Sal), dissolved inorganic carbon (DIC), total alkalinity (TA), aragonite saturation (Ω_A), nutrient concentrations summarized by their first principal component (Nut), chlorophyll-a ($Chla$), seawater pH (pH_{sw}), partial pressure of carbon dioxide (P_{CO_2}), and net ecosystem calcification (NEC). The underlying structure is adapted from Courtney et al. (2017), following the acyclic construction of Aglietti et al. (2020). The structural equations are obtained by performing linear regression on the available standardized observations from Andersson & Bates (2018), using the 95 records supplied with the implementation. The graph structure underlying the Bermuda task has two root variables, $Light$ and Nut. The parent sets of

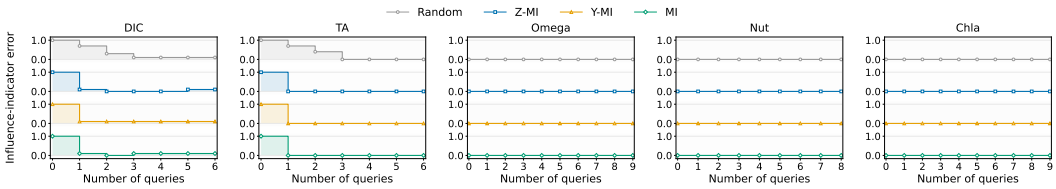

Figure 6: Influence-indicator mistake rates versus the number of queries in BERMUDA.

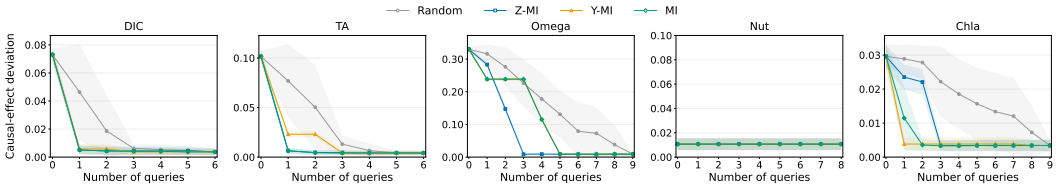

Figure 7: Causal-effect deviations versus the number of queries in BERMUDA.

the remaining variables are

$$\begin{aligned}
\mathbf{Pa}_{Tem} &= \{Light\},\\
\mathbf{Pa}_{Sal} &= \{Tem\},\\
\mathbf{Pa}_{DIC} = \mathbf{Pa}_{TA} &= \{Sal\},\\
\mathbf{Pa}_{P_{CO_2}} = \mathbf{Pa}_{pH_{sw}} = \mathbf{Pa}_{\Omega_A} &= \{DIC, Sal, TA, Tem\},\\
\mathbf{Pa}_{Chla} &= \{Light, Nut, Tem\},\\
\mathbf{Pa}_{NEC} &= \{P_{CO_2}, Chla, Light, Nut, \Omega_A, Tem, pH_{sw}\}.
\end{aligned}$$

Accordingly, the common-cause structures for the five actionable variables are

$$\begin{aligned}
\mathcal{C}_{DIC} = \mathcal{C}_{TA} &= \{Light, Tem, Sal\},\\
\mathcal{C}_{\Omega_A} &= \{Light, Tem, Sal, DIC, TA\},\\
\mathcal{C}_{Nut} &= \emptyset,\\
\mathcal{C}_{Chla} &= \{Light, Tem, Nut\}.
\end{aligned}$$

To ensure compatibility with our theoretical framework, we discretize each variable into three equal-width bins on $[-0.15, 0.15]$, indexed by $\{0, 1, 2\}$ and represented by -0.1, 0, and 0.1, respectively. We take NEC as the outcome variable and consider DIC, TA, Ω_A, Nut, and $Chla$ as the single actionable variable Z. The action space is $\Delta_Z = \{0, 1, 2, \perp\}$, where $\perp$ denotes inaction. The desired set is $\mathcal{S} = \{2\}$, corresponding to the highest calcification state. With an odds threshold of one, the objective is to determine which actions make the desired outcome at least as probable as its complement. We use 100,000 observational samples in each of ten independent repetitions.

E ADDITIONAL EXPERIMENTS

Figs. 6 and 7 show the full results for all the actionable variables and query strategies in BERMUDA. For DIC and TA, the indicator mistake rates become very small before the corresponding causal-effect estimates fully converge. For Ω_A, Nut, and $Chla$, the indicators are correctly determined before any queries, while the causal-effect estimates still deviate from their true values. Notably, this distinction is observed under both random and MI-based strategies, rather than being specific to the proposed strategy. These results further support our theoretical insight. Limited common-cause knowledge can provide useful decision guidance before the corresponding causal effects are accurately estimated. The four query selection strategies are also compared. In most cases, the MI-based strategies reduce indicator mistakes with fewer queries than RANDOM, with MI performing comparably to the two MI baselines. This validates the effectiveness of active querying in acquiring relevant common-cause knowledge when the influence indicators need to be determined.