ON MEASURING INFLUENCE IN AVOIDING UNDESIRED FUTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

When a predictive model anticipates an undesired future event, a question arises: what can we do to avoid it? The key to resolving this forward-looking challenge lies in determining the right variables that influence the result, moving beyond statistical correlations typically exploited for prediction. In this paper, we introduce a novel framework for evaluating the influence of alterable variables in successfully avoiding the undesired future. We quantify influence as the degree to which the probability of success can be increased by altering variables based on the principle of maximum expected utility. A crucial insight from our analysis is that the most influential variables may not necessarily be those with inherently strong causal effects on the future event. In fact, it can be highly beneficial to alter a weak causal ancestor, or even a variable that is not a causal ancestor at all. Furthermore, to overcome the practical challenges of exact computation, we provide a Monte-Carlo method for efficiently assessing influence using observational data. Experiments demonstrate the empirical performance of the proposed framework.

1 INTRODUCTION

When an intelligent machine receives a warning from a powerful predictive model anticipating that an undesired event is going to happen, an important question naturally arises: what can be done to avoid this potential future? This is known as the *avoiding undesired future* (AUF) problem (Zhou, 2022), sparking a transition from passively predicting results to proactively *influencing* them.

Addressing the AUF problem requires determining the variables that can be properly altered to shape a more desirable future. While statistically correlated variables are effectively exploited by modern machine learning (ML) techniques for predicting target variables (Jumper et al., 2021; Achiam et al., 2023; Price et al., 2025), these correlations are often unreliable for influencing the future target. For instance, although ice cream sales and drowning incidents are highly correlated in the summer, suppressing ice cream sales would obviously not prevent drownings, as their superficial correlation arises from a common cause: hot weather. This implies that a general understanding of the underlying mechanisms connecting variables would be essential for settling the AUF problem.

To this end, an intuitive strategy is to exploit causal variables of the target. Rich tools for discovering causal relations have been developed in the literature (Pearl, 2009; Peters et al., 2017). Nevertheless, the fact that a variable is a cause of the target variable does not imply that altering it will be influential. For example, while a city's reliance on public transportation might be a cause of lengthy commute times, a policy encouraging the use of private cars could fail to save time due to offsetting effects: the positive impact on shortening commute times obtained via private cars could be neutralized by the negative aspect, such as the worsening traffic congestion caused by much more cars on the road. This seems to suggest focusing our attention on variables with non-negligible average causal effects. However, this strategy is also insufficient. As illustrated by the simple case of two alterable variables in Figure 1, it can be highly influential to alter a variable with a negligible causal effect. Therefore, a more principled way is needed to properly address the AUF problem.

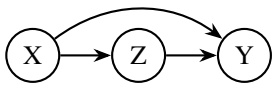


Figure 1: *To do, or not to do, that is the question*: whether a variable should be altered to *influence* the eventuality? Here, X and Z are both alterable variables, and Y is target variable. Let X be a Bernoulli variable, with $Z := 1 - X$ and $Y := \min(X, Z)$. Clearly, while the average causal effect of X on Y is 0, it is indeed beneficial to alter X (and Z). More details are in Example 3.

In this paper, we introduce a novel framework for measuring the influence of alterable variables in successfully avoiding the undesired future. We begin by outlining several natural and intuitive considerations that a measure of influence in AUF scenarios should incorporate. Then, we formulate a new quantity, termed *influence power*, defined as the degree to which the probability of success can be increased through alterations. Grounded in the principle of maximum expected utility, this quantity systematically accounts for the *alterability*, *naturality*, and *desirability* of variables throughout the decision process. In this way, the influence power offers a holistic assessment of the consequence of alteration, capturing both its explicit and implicit impacts on the future target.

Next, we leverage influence power to methodically investigate the relationship between influential variables and those with causal links to the future target. Our investigation indicates a subtle yet important distinction: influential variables are not simply a subset of causal ancestors, and vice versa. Although influence often arises from causal effects, we show that influential variables are not necessarily those with inherent causal effects on the future target. In fact, it can be highly beneficial to alter a causal ancestor with negligible effects, or even a variable that is not a causal ancestor at all. Another important observation is that, not all alterable variables can be safely altered, as for certain variables, any alteration is counterproductive. This insight crystallizes the fundamental question for any agent facing an undesired future: *To do, or not to do?* Our framework rests on a principled quantity for measuring influence in AUF, thereby providing a rigorous way to answer this question.

Finally, we address the practical computation of influence power. We address the challenges inherent in its exact computation and then present a Monte-Carlo-based approximation method to efficiently assess it using observational data. Notably, our method circumvents the need for causal information of structural equations, and tends to remain effective even when the probability terms within our quantity are not accurately approximated. Our empirical results demonstrate the effectiveness and efficiency of the proposed method in measuring influence in AUF under limited quality of approximation.

2 PRELIMINARY

Notation. We represent each random variable with a capital letter (V), and its realized value with the lowercase letter (v). We use bold capital letters ($\mathbf{V}$) to denote a set of random variables with their realized values denoted by bold lowercase letters ($\mathbf{v}$). Let $G = (\mathbf{V}, \mathbf{E})$ denote a directed graph with nodes $\mathbf{V}$ and edges $\mathbf{E}$. We say that a variable X is causally linked to another variable Y, or X is a causal ancestor of Y, if there exists a directed path from X to Y in G. When X is binary, its causal strength can be quantified by the *average causal effect* (ACE) (Holland, 1988; Pearl, 2009), defined as $\tau(X, Y) \coloneqq \mathbb{E}(Y|do(X = 1) - \mathbb{E}(Y|do(X = 0)))$, where $\mathbb{E}(Y|do(X = x))$ denotes the expectation of Y when X is set to the value x. We say that a causal ancestor X of Y is weak when the average causal effect of X on Y is nearly negligible. Let Δ_X denote the feasible domain of alteration for a variable X. If $\Delta_X \neq \emptyset$, we call X an alterable variable; otherwise, X is unalterable.

Structural Causal Models. We use the language of *structural causal models* (SCMs) (Pearl, 2009) to describe the physical mechanisms governing the generating process of variables. An SCM is a tuple $\mathcal{M} = \langle \mathbf{V}, \mathbf{N}, F, P(\mathbf{N}) \rangle$, where $\mathbf{V} = (V_1, \ldots, V_d)$ is a set of observable variables, $\mathbf{N} = (N_1, \ldots, N_d)$ is a set of independent background noises destributed according to $P(\mathbf{N})$, and F is a set of deterministic functions f_i for each $V_i \in \mathbf{V}$ such that $V_i \coloneqq f_i(\mathrm{PA}_i, N_i)$ with $\mathrm{PA}_i \subseteq \mathbf{V}$. For a variable $V_i \in \mathbf{V}$, if $\Delta_{V_i} \neq \emptyset$, we use the notation $V_i \overset{a}{=} v_i$ to indicate that V_i can be altered to $v_i \in \Delta_{V_i}$. This alteration is formally represented by replacing the function for V_i in $\mathcal{M}$ with the assignment $V_i \coloneqq v_i$. The resulting SCM is denoted as $\mathcal{M}_{V_i}$. The distribution of variables $\mathbf{W} \subseteq \mathbf{V}$ in $\mathcal{M}_{V_i}$ is then denoted as $P(\mathbf{W}|V_i \overset{a}{=} v_i)$, i.e., the distribution of $\mathbf{W}$ given that V_i is altered to v_i.

Problem Definition. We consider a scenario where observational data $\mathcal{D}$ is drawn from a distribution P, induced by an underlying SCM $\mathcal{M}$ over ordered variables $(\mathbf{V}, Y)$, with Y being the final variable in the sequence. We assume for simplicity that variables are discrete and causally ordered. Furthermore, each variable $V_i \in \mathbf{V}$ is accompanied by an alterable domain Δ_{V_i}, which is known beforehand. Throughout this paper, we use Y to denote the target variable and $\mathcal{S}$ to denote its desired region. Now, a new observation $\mathbf{x}$ for a subset of variables $\mathbf{X} \subseteq \mathbf{V}$ appears, a predictive model h outputs a prediction $\hat{y} = h(\mathbf{x})$, and a warning is triggered if $\hat{y}$ falls outside a predefined desired region $\mathcal{S}$. Then, the main goal of the AUF problem is to make feasible alterations to the subsequent variables before the target variable Y is finalized, ensuring that Y falls into the desired region $\mathcal{S}$ as much as possible.

3 INFLUENCE POWER

3.1 MOTIVATION

We motivate the considerations that a measure of influence in AUF should incorporate by describing the strategies and limitations of existing approaches for addressing the AUF problem.

A primary strategy is to find a feasible alteration that directly maximizes the probability of Y falling within the desired region $\mathcal{S}$ (Qin et al., 2023). This straightforward strategy is expressed as:[1]

$$(Z^*, z^*) = \arg\max_{Z \in \mathbf{Z}, z \in \Delta_Z} P(Y \in \mathcal{S} | Z \overset{a}{=} z), \tag{1}$$

where $\mathbf{Z} \subseteq \mathbf{V}$ denotes the set of alterable variables. This approach is intuitive and can indeed achieve a better target in many cases, but it overlooks several important considerations. Specifically, Equation (1) only accounts for the straightforward effect of altering a single variable at a time, presuming a "static" future where subsequent variables unfold naturally. Thus, an immediate consequence is that it ignores how multiple variables might combine their effects. A very simple example illustrates this issue. Imagine two binary variables, Z_1 and Z_2, both of which naturally take the value 0 with near certainty, and let $Y \coloneqq Z_1 \wedge Z_2$. Clearly, altering either variable alone is ineffective. It's only by altering both variables together that we can achieve $Y = 1$. Consequently, when judging the impact of an alteration in AUF scenarios, not only the feasible domain of the alteration itself but also the *alterability* of other variables should be considered.

Given the insight from the example above, the next logical step would be to propose the joint alteration of all alterable variables as a solution. This joint strategy has been adopted in previous work (Qin et al., 2025; Du et al., 2025) with the following formulation:

$$(z_1^*, \ldots, z_m^*) = \arg\max_{z_1 \in \Delta_{Z_1}, \ldots, z_m \in \Delta_{Z_m}} P(Y \in \mathcal{S} | Z_1 \overset{a}{=} z_1, \ldots, Z_m \overset{a}{=} z_m), \tag{2}$$

where $m = |\mathbf{Z}|$ is the number of alterable variables. While the described strategy works for the case of $Y \coloneqq Z_1 \wedge Z_2$, it overlooks an important fact: it's often unnecessary to alter all variables. For instance, while both light and water are crucial factors for crop growth, if sunlight is naturally abundant, adding artificial light will have no impact on yield and instead leads to unnecessary costs. Therefore, when judging the impact of altering a variable in AUF scenarios, we need to consider its *naturality*, i.e., whether it is already in a favorable state naturally. Moreover, as we shall see in what follows, certain variables may not only be unnecessary to alter, but could even be counterproductive no matter how they are altered. Thus, a better, more principled approach is required to determine which alterable variables should be altered instead of indiscriminately altering all of them.

3.2 FORMULATION

In this subsection, we formulate a new quantity that measures whether an alterable variable are worth altering in order to influence the future target. To holistically account for the alterability and naturality of variables, as well as the desirability of the target variable in the decision process, our formulation requires a principled way to envision future possibilities after an alteration. The Bellman equation (Bellman, 1957) provides the conceptual foundation for this purpose, but its standard formulation is not immediately applicable to our context. This is because the classical framework is usually built upon a prespecified separation between state and control variables. In the AUF problem, however, every V_i in the sequence of variables $\mathbf{V} = (V_1, \ldots, V_d)$ has a dual role: it could be proactively manipulated through alteration or be passively observed as it unfolds naturally.

Drawing inspiration from the Bellman equation and grounding our proposal in the principle of maximum expected utility (Russell & Norvig, 2020), we recursively define the *maximum expected probability* (MEP) of avoiding the undesired future after an alteration or observation. Specifically, if $k = d$, the MEP after *altering* V_k to v_k, denoted as $\mathcal{P}(Y \in \mathcal{S} | V_k \overset{a}{=} v_k, \ldots)$, simply equals to the *AUF probability* $P(Y \in \mathcal{S} | V_k \overset{a}{=} v_k, \ldots)$, where "$\ldots$" abbreviates any form of alterations and observations that happened before V_k. For $0 < k < d$, the MEP after *altering* V_k to v_k is given by

$$\begin{aligned}\mathcal{P}(Y \in \mathcal{S} | V_k \overset{a}{=} v_k, \ldots) \coloneqq \max \Big\{ & \max_{v_{k+1} \in \Delta_{V_{k+1}}} \mathcal{P}(Y \in \mathcal{S} | V_{k+1} \overset{a}{=} v_{k+1}, V_k \overset{a}{=} v_k, \ldots), \\ & \mathbb{E}_{v_{k+1} \sim P(V_{k+1} | V_k \overset{a}{=} v_k, \ldots)} \mathcal{P}(Y \in \mathcal{S} | V_{k+1} \overset{o}{=} v_{k+1}, V_k \overset{a}{=} v_k, \ldots) \Big\},\end{aligned} \tag{3}$$

[1] For clarity, the observation of $\mathbf{X} = \mathbf{x}$ is omitted from the condition of the probability $P(Y \in \mathcal{S} | Z \overset{a}{=} z)$.

where $\mathcal{P}(Y \in \mathcal{S}|V_{k+1} \overset{o}{=} v_{k+1}, V_k \overset{a}{=} v_k, \ldots)$ is interpreted as the MEP after the observation of $V_{k+1} \overset{o}{=} v_{k+1}$ and the alteration of $V_k \overset{a}{=} v_k$. Similarly, if $j = d$, the MEP after *observing* V_j as v_j, denoted as $\mathcal{P}(Y \in \mathcal{S}|V_j \overset{o}{=} v_j, \ldots)$, simply equals to the AUF probability $P(Y \in \mathcal{S}|V_j \overset{o}{=} v_j, \ldots)$. For $0 < j < d$, the MEP after *observing* V_j as v_j is similarly given by

$$\mathcal{P}(Y \in \mathcal{S}|V_j \overset{o}{=} v_j, \ldots) \coloneqq \max \Big\{ \max_{v_{j+1} \in \Delta_{V_{j+1}}} \mathcal{P}(Y \in \mathcal{S}|V_{j+1} \overset{a}{=} v_{j+1}, V_j \overset{o}{=} v_j, \ldots), \quad (4)$$
$$\mathbb{E}_{v_{j+1} \sim P(V_{j+1}|V_j \overset{o}{=} v_j, \ldots)} \mathcal{P}(Y \in \mathcal{S}|V_{j+1} \overset{o}{=} v_{j+1}, V_j \overset{o}{=} v_j, \ldots) \Big\}.$$

Based on the above recursive definition of MEP, we formulate a quantity called the *influence power*, indicating the ability of an alterable variable to influence the future target.

Definition 1 (Influence Power)**.** *For any $V_i \in \mathbf{V}$, the influence power of V_i on Y is defined as*

$$\dot{p}(V_i, Y) \coloneqq \max_{v_i \in \Delta_{V_i}} \mathcal{P}(Y \in \mathcal{S}|V_i \overset{a}{=} v_i) - \mathbb{E}_{v_i \sim P(V_i)} \mathcal{P}(Y \in \mathcal{S}|V_i \overset{o}{=} v_i).$$

Remark. The influence power of V_i on Y represents the maximum increase in the MEP that can be achieved by optimally altering V_i, compared to the expected MEP when V_i is observed naturally. Consequently, a positive influence power indicates that an alteration is beneficial, while a zero or negative influence power suggests that it is unnecessary or even harmful. By definition, the influence power is bounded within the range of $[-1, 1]$. It is noteworthy that this concept can be easily extended to a conditional form. For example, given the observation that a set of variables $\mathbf{X}$ takes the value $\mathbf{x}$, the conditional influence power of V_i on Y is given by $\dot{p}(V_i, Y|\mathbf{X} \overset{o}{=} \mathbf{x}) \coloneqq \max_{v_i \in \Delta_{V_i}} \mathcal{P}(Y \in \mathcal{S}|V_i \overset{a}{=} v_i, \mathbf{X} \overset{o}{=} \mathbf{x}) - \mathbb{E}_{v_i \sim P(V_i|\mathbf{X} \overset{o}{=} \mathbf{x})} \mathcal{P}(Y \in \mathcal{S}|V_i \overset{o}{=} v_i, \mathbf{X} \overset{o}{=} \mathbf{x})$. Furthermore, as Definition 1 recursively follows the principle of maximum expected utility, the influence power can be interpreted as a variant of the Bellman equation.

We end this subsection by highlighting an intriguing connection between Definition 1 and Equation (1). Consider an extremely simple setting involving three binary variables: V_1, V_2, and Y, where both V_1 and V_2 are alterable. It has been informed by an oracle that the deterministic function defining the target variable Y depends solely on V_1 and not V_2, i.e., $Y \coloneqq f(V_1)$. Given this, it is evident that the solution to Equation (1) is V_1 when the following condition holds:

$$\max_{v_1 \in \Delta_{V_1}} P(Y \in \mathcal{S}|V_1 \overset{a}{=} v_1) > P(Y \in \mathcal{S}). \quad (5)$$

Meanwhile, in this case, it is clear that the influence power of V_1 on Y can be simplified to $\dot{p}(V_1, Y) = \max_{v_1} P(Y \in \mathcal{S}|V_1 \overset{a}{=} v_1) - P(Y \in \mathcal{S})$. This implies that the solution of Equation (1) is V_1 when $\dot{p}(V_1, Y) > 0$. In other words, Definition 1 and Equation (1) agree on determining whether V_1 should be altered in this simple setting. Interestingly, by further assuming $\mathcal{S} = \{1\}$ and $\Delta_{V_1} = \{0, 1\}$, and using the identity $2 \cdot \max(a, b) = a + b + |a - b|$, we can deduce that Equation (5) is equivalent to $|\tau(V_1, Y)| = |P(Y = 1|V_1 \overset{a}{=} 1) - P(Y = 1|V_1 \overset{a}{=} 0)| > \gamma$, where the threshold $\gamma = 2P(Y \in \mathcal{S}) - P(Y = 1|V_1 \overset{a}{=} 0) - P(Y = 1|V_1 \overset{a}{=} 1)$, stating that the average causal effect of V_1 on Y, denoted as $\tau(V_1, Y)$, is non-negligible compared with c. Consequently, in this case, the variable V_1 is worth altering if and only if its average causal effect on Y is absolutely non-negligible.

While the simple case above suggests that Definition 1 and the baseline approach from Equation (1) both favor altering variables with strong causal effects, this view is incomprehensive. In the following, we will show that the relationship between influential and causal variables is far more subtle.

3.3 Investigation

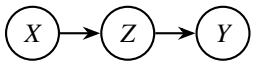
We present a spectrum of configurations with concrete examples to investigate the relationship between influential variables and causal ancestors, offering valuable insights into the AUF problem.

Causal Ancestors with No Influence. A variable can be a causal ancestor of the target yet have no influence. This situation can arise trivially if a cause is unalterable. For instance, the past cannot be changed, as anything that has already occurred is immutable. Also, a variable such as a person's age is unalterable. Beyond this, an alterable variable can also be non-influential. This can occur when its causal effect is negligible, such as in the example discussed in Section 1 where the positive and negative impacts of the variable on the target balanced out, resulting in offsetting effects.

Strong Causal Ancestors with No Influence. A causal ancestor with a strong average causal effect can also have no influence on the target variable.

Example 1. *Consider the following structural equations with the corresponding causal graph:*

$$X := N_X,$$
$$Z := X \cdot N_Z + (1 - X) \cdot (1 - N_Z),$$
$$Y := Z \cdot N_Y + (1 - Z) \cdot (1 - N_Y),$$



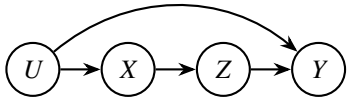
where $N_X, N_Z, N_Y \overset{iid}{\sim} \text{Bern}(0.9)$. *Let* X *and* Z *be alterable variables, let* $\Delta_X = \{0, 1\}$ *and* $\Delta_Z = \{0, 1\}$ *be the feasible domains of alteration, and let the desired region for* Y *be* $\mathcal{S} = \{1\}$.

In this example, the average causal effect of X on Y is also non-negligible. Concretely, $\tau(X, Y) = P(Y = 1|X \overset{a}{=} 1) - P(Y = 1|X \overset{a}{=} 0) = 0.82 - 0.18 = 0.64$. In contrast, the influence power of X on Y can be simplified to $\dot{p}(X, Y) = \max_{x \in \Delta_X} \max_{z \in \Delta_Z} P(Y = 1|Z \overset{a}{=} z, X \overset{a}{=} x) - \mathbb{E}_{x \sim P(X)} \max_{z \in \Delta_Z} P(Y = 1|Z \overset{a}{=} z, X \overset{o}{=} x) = 0.9 - 0.9 = 0$. Again, this indicates that the alteration on X is unnecessary, as it makes no difference on the probability of $Y = 1$, given that a rational agent will always alter Z to 1 to maximize the probability of $Y = 1$. In short, it is useless to alter X in Example 1 because X is shielded by the alterability of Z.

Strong Causal Ancestors with Negative Influence. Perhaps the most counter-intuitive case is when a variable with a non-negligible ACE can be not only useless to alter, but also detrimental.

Example 2. *Consider the following structural equations with the corresponding causal graph:*

$$U := N_U,$$
$$X := U \cdot N_X + (1 - U) \cdot (1 - N_X),$$
$$Z := X \cdot N_Z + (1 - X) \cdot (1 - N_Z),$$
$$Y := Z \cdot (1 - U) + (1 - Z) \cdot N_Y,$$



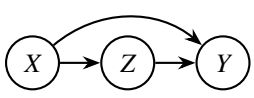
where $N_U, \sim \text{Bern}(0.5)$, $N_X, N_Z \overset{iid}{\sim} \text{Bern}(0.9)$, *and* $N_Y \sim \text{Bern}(0.4)$. *Let* X *and* Z *be alterable variables with* $\Delta_X = \{0, 1\}$ *and* $\Delta_Z = \{0, 1\}$, *and let the desired region be* $\mathcal{S} = \{1\}$.

In this example, we have the average causal effect of X on Y: $\tau(X, Y) = P(Y = 1|X \overset{a}{=} 1) - P(Y = 1|X \overset{a}{=} 0) = 0.49 - 0.41 = 0.08$. However, the influence power of X on Y is negative: $\dot{p}(X, Y) = \max_{x \in \Delta_X} \mathcal{P}(Y = 1|X \overset{a}{=} x) - \mathbb{E}_{x \sim P(X)} \mathcal{P}(Y = 1|X \overset{o}{=} x) = 0.5 - 0.65 = -0.15$. This indicates that the MEP after altering X is always less than the expected MEP after observing X. In other words, any alteration on X is counterproductive regardless of how X is altered. Intuitively, this is because observing X somehow reflects information about U, and this reflection would help the rationality of alteration on Z within the computation of $\dot{p}(X, Y)$. Hence, although the alteration of X can lead to a straightforward effect on improving Y, its positive impact is overturned by the negative impact on the alteration of the subsequent variable, making it detrimental to alter X. Moreover, based on the intuition that observing X reflects U, it is interesting to evaluate the conditional influence power of X on Y given the observation of U. Specifically, suppose that U has been observed to be 1, we have $\dot{p}(X, Y|U \overset{o}{=} 1) = 0$, stating that altering X has no influence on Y. Thus, it is unnecessary to alter X given the observation of U. This verifies the intuition above. We further point out that, when U has been observed, Example 2 becomes reduces to a similar case of Example 1 where X makes no difference because it is shileded by the alterability of Z. In practice, it could be plausible that directly observing a variable, such as U in Example 2, is time-consuming and expensive. When it is unable to timely observe the value of U, the conditional influence power $\dot{p}(X, Y|U \overset{o}{=} 1) = 0$ could not be evaluated, and thus the influence power $\dot{p}(X, Y)$ remains instructive for the agent.

Weak Causal Ancestors with Positive Influence. In direct contrast to the preceding case, a variable with a negligible average causal effect can have a positive influence power.

Example 3. *Consider the following structural equations with the corresponding causal graph:*

$$X = N_X,$$
$$Z = (1 - X) \cdot N_Z,$$
$$Y = X \cdot Z \cdot N_Y,$$



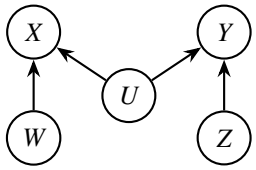
where $N_X, N_Z, N_Y \overset{iid}{\sim} \text{Bern}(0.9)$. *Let* X *and* Z *be alterable variables with* $\Delta_X = \{0, 1\}$ *and* $\Delta_Z = \{0, 1\}$, *and let the desired region for* Y *be* $\mathcal{S} = \{1\}$.

In this example, while X is a causal ancestor of Y, the average causal effect of X on Y is negligible: $\tau(X, Y) = 0$. Meanwhile, the influence power of X on Y is positive: $\dot{p}(X, Y) = 0.9 - 0.81 = 0.09$. Intuitively, the influence of X arises from its collaboration with the alterability of Z. Only by considering to alter Z, the implicit impact of X on Y can be elicited. The implicit impact of X on Y can only be elicited when the potential alteration of Z is also considered. Influence power captures this implicit gain, whereas the $\tau(X, Y)$ alone would dismiss Z as negligible.

Non-Causal Variables with Positive Influence. Finally, we present an example showing that a variable can have positive influence power despite it is not a causal ancestor of the target.

Example 4. *Consider the following structural equations with the corresponding causal graph:*

$$
\begin{aligned}
U &= N_U, \\
W &= N_W, \\
X &= U \cdot W \cdot (1 - N_X), \\
Z &= N_Z, \\
Y &= Z \cdot (1 - U) + (1 - Z) \cdot N_Y,
\end{aligned}
$$



where $N_U \sim \text{Bern}(0.5)$, $N_W, N_X, N_Z \overset{iid}{\sim} \text{Bern}(0.1)$, *and* $N_Y \sim \text{Bern}(0.4)$. *Let* W, X, *and* Z *be alterable variables with* $\Delta_W = \Delta_X = \Delta_Z = \{0, 1\}$, *and let the desired region be* $\mathcal{S} = \{1\}$.

In this example, W is not a causal ancestor of Y, and thus its average causal effect is zero: $\tau(W, Y) = 0$. However, the influence power of W on Y is positive: $\dot{p}(W, Y) = 0.68 - 0.518 = 0.162$. This indicates that altering W can significantly improve the MEP of $Y = 1$. Intuitively, this positive influence arises because altering W helps X to carry information about U. This information, in turn, enables a more rational subsequent alteration of Z, which ultimately impacts Y. Influence power successfully captures this implicit, indirect benefit, highlighting that even non-causal variables can be crucial for AUF. For completeness, we can also evaluate the influence power of X on Y: $\dot{p}(X, Y) = 0.68 - 0.518 = -0.162$. This is similar to the case in Example 2, where altering X is counterproductive as its explicit positive impact is overturned by its negative impact on the alteration of Z. If the variable U were observed, the influence of both X and W would be nullified. For instance, given $U \overset{o}{=} 1$, we have $\dot{p}(X, Y | U \overset{o}{=} 1) = 0$ and $\dot{p}(W, Y | U \overset{o}{=} 1) = 0$. Nonetheless, as we have discussed, it is often time-consuming or unaffordable to promptly observe U in practice. Therefore, the quantities $\dot{p}(W, Y)$ and $\dot{p}(X, Y)$ remain instructive and valuable for the AUF problem.

4 Assessing Influence Power

In this section, we provide an efficient method to assess the influence power defined in Definition 1. While influence power is a principled quantity for measuring the influence of alterable variables, its investigation in Section 3.3 requires two conditions that are often impractical: an exhaustive computation of the MEP terms and knowledge of the underlying structural equations. In what follows, we address each of these obstacles. We then discuss how our method can still yield an informative indicator of influence even when the MEP cannot be precisely estimated due to practical constraints.

4.1 Monte-Carlo Approximation

The recursive enumeration of MEP for all possible alterations can be computationally prohibitive when the number of alterable variables is large. To mitigate this, we interpret the computation of MEP as a single-player non-deterministic game and approximate it based on the Monte-Carlo tree search UCT (Upper Confidence Tree) introduced by Kocsis & Szepesvári (2006).

Specifically, a *search tree* employing Monte-Carlo simulations is constructed incrementally. Each *node* in the tree represents a *state* defined by a sequence of alterations and observations made so far, associated with the next variable to be considered. Every iteration begins at the root node N_0 (associated with a pre-specified variable $V_i \in \mathbf{V}$), proceeds to its children (associated with V_{i+1}), and continues until reaching a terminal state (associated with the target variable Y). Each *edge* in the tree represents a *choice* that can be made from the node, i.e., either an alteration or an observation on the associated variable. The overall construction consists of four steps, iterated until time has expired: (1) *Selection*: starting from the root node, recursively select an edge to child nodes according to the

UCT policy until reaching a leaf node; (2) *Expansion*: if the leaf node corresponds to a non-terminal state, expand it by randomly adding one child node corresponding to possible choices; (3) *Playout*: from the newly added node, execute a random sequence of choices until reaching a terminal state, and compute the AUF probability at that terminal state; (4) *Backpropagation*: propagate the computed AUF probability back up the tree, updating the statistics of each node along the path. During each iteration, the UCT criterion is used at a node N to select the next edge to traverse:

$$c_N^* = \arg\max_{c \in \Delta_N^+} \left\{ \hat{p}_{N,c} + \alpha \cdot \sqrt{\tfrac{\ln t_N}{t_{N,c}}} \right\}, \tag{6}$$

where $\Delta_N^+ = \Delta_N \cup \emptyset$ is the set of choices at node N (comprising feasible alterations on the variable associated with N, denoted by Δ_N, and the option to make an observation, denoted by $\emptyset$), $\hat{p}_{N,c}$ is the average AUF probability obtained after taking choice c at node N, α is a parameter used to balance between exploration and exploitation (Auer et al., 2002), t_N is the number of times node N has been selected, and $t_{N,c}$ is the number of times choice c has been selected at node N.

After the construction of search tree, the MEP terms in the influence power of V_i on Y are approximated as the average AUF probability for each choice at the root node N_0 of search tree. Concretely, we have $\mathcal{P}(Y \in \mathcal{S} | V_i \overset{a}{=} c) \approx \hat{p}_{N_0,c}$ for each $c \in \Delta_{N_0}$, and $\mathbb{E}_{v_i \sim P(V_i)} \mathcal{P}(Y \in \mathcal{S} | V_i \overset{o}{=} v_i) \approx \hat{p}_{N_0,\emptyset}$. Hence, according to Definition 1, the influence power of V_i on Y is approximated as

$$\dot{p}(V_i, Y) \approx \max_{c \in \Delta_{N_0}} \hat{p}_{N_0,c} - \hat{p}_{N_0,\emptyset}. \tag{7}$$

The quality of this approximation improves over time, as UCT is guaranteed to converge to the best choice given enough iterations. Moreover, the described procedure is an *anytime* algorithm, capable of producing an approximate influence power at any point during its computation. We refer the reader to Browne et al. (2012) for further details.

4.2 AUF Probability Estimation

Although the Monte-Carlo procedure described above can effectively approximate the influence power, it still relies on the AUF probability when a terminal state is reached during simulations, whose ground-truth value is dictated by the underlying SCM. For situations where the structural equations are unknown, we present an expression for estimating the AUF probability from observational data.

Specifically, we express the joint probability of the ordered variables $(\mathbf{V}, Y)$ as:

$$P(\mathbf{V}, Y) = P(V_1, \ldots, V_d, Y) = P(Y|\mathbf{V}) \prod_{i=1}^{d} P(V_i | V_1, \ldots, V_{i-1}), \tag{8}$$

where the conditional probabilities $P(Y|\mathbf{V})$ and $P(V_i|V_1, \ldots, V_{i-1})$ can be estimated from observational data $\mathcal{D} = \{(\mathbf{v}^j, y^j)\}_{j=1}^n$ using standard ML techniques. Denote by $\mathbf{A}$ the variables in $\mathbf{V}$ that are altered, we express the joint probability of $(\mathbf{V}, Y)$ given the alteration of $\mathbf{A}$ as follows:

$$P(\mathbf{V}, Y | \hat{\mathbf{A}}) = P(Y|\mathbf{V}) \prod_{V_i \in \mathbf{A}} \delta(V_i) \prod_{V_i \in \mathbf{V} \setminus \mathbf{A}} P(V_i | V_1, \ldots, V_{i-1}). \tag{9}$$

Then, denote by $\mathbf{O}$ the variables in $\mathbf{V}$ that are observed, the AUF probability given the alteration of $\mathbf{A}$ and the observation of $\mathbf{O}$ is expressed as:

$$\begin{aligned} P(Y \in \mathcal{S} | \hat{\mathbf{A}}, \mathbf{O}) &= \frac{P(Y \in \mathcal{S}, \mathbf{O} | \hat{\mathbf{A}})}{P(\mathbf{O} | \hat{\mathbf{A}})} = \frac{\sum_{\mathbf{V} \setminus \mathbf{O}} P(Y \in \mathcal{S}, \mathbf{V} | \hat{\mathbf{A}})}{\sum_{\mathbf{V} \setminus \mathbf{O}} P(\mathbf{V} | \hat{\mathbf{A}})} \\ &= \frac{\sum_{\mathbf{V} \setminus \mathbf{O}} P(Y \in \mathcal{S} | \mathbf{V}) \prod_{V_i \in \mathbf{A}} \delta(V_i) \prod_{V_i \in \mathbf{V} \setminus \mathbf{A}} P(V_i | V_1, \ldots, V_{i-1})}{\sum_{\mathbf{V} \setminus \mathbf{O}} \prod_{V_i \in \mathbf{A}} \delta(V_i) \prod_{V_i \in \mathbf{V} \setminus \mathbf{A}} P(V_i | V_1, \ldots, V_{i-1})}, \end{aligned} \tag{10}$$

which is a generic expression of the AUF probability given any alterations and observations. It can be estimated from observational data $\mathcal{D}$ and then plugged into the Monte-Carlo procedure described above to approximate the influence power. The following proposition demonstrates the consistency of Equation (10) by leveraging the manipulation theorem in Spirtes et al. (2000).

Proposition 1. *Assume causal sufficiency, i.e., the joint distribution $P(\mathbf{V}, Y)$ is induced by an acyclic SCM $\mathcal{M}$ with mutually independent background noises. Then, the expression in Equation* (9) *is consistent to the joint probability dictated by the SCM $\mathcal{M}_{\mathbf{A}}$ where variables $\mathbf{A}$ are altered. Furthermore, the expression in Equation* (10) *is consistent to the AUF probability dictated by the SCM $\mathcal{M}_{\mathbf{A}}$ where variables $\mathbf{A}$ are altered and variables $\mathbf{O}$ are observed.*

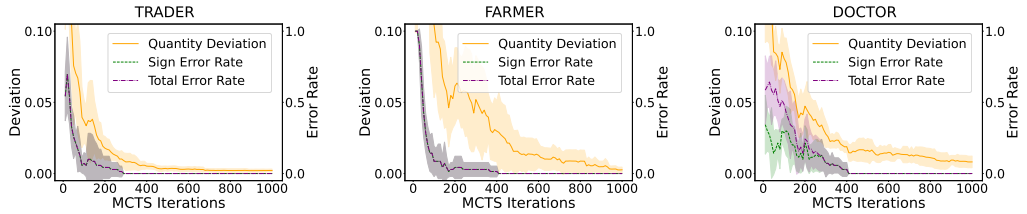


Figure 2: Convergence of the approximation of influence power and error rates (%) versus the number of MCTS iterations. The deviation of the approximated influence power to the exact value of influence power continues to decrease after the convergence of error rates in all cases.

TASK	MAX-ONE	MAX-ALL	OURS ($T = 10$)	OURS ($T = 50$)	OURS ($T = 100$)	OURS ($T = 250$)
TRADER	57.76 ± 9.15	57.74 ± 9.85	58.74 ± 9.38	60.74 ± 7.42	60.77 ± 7.43	**60.77 ± 7.43**
FARMER	48.91 ± 13.37	59.69 ± 19.79	71.57 ± 17.28	73.20 ± 14.08	74.22 ± 13.61	**74.22 ± 13.61**
DOCTOR	49.73 ± 5.57	49.83 ± 5.56	53.79 ± 8.61	54.54 ± 6.77	58.81 ± 5.85	**62.19 ± 6.39**

Table 1: Comparison of probability of success (%) in three tasks.

4.3 DISCUSSION

An efficient way to approximate influence power from observational data has been presented by combining the Monte-Carlo approximation with the AUF probability estimation, and the accuracy of this approximation requires enough iterations of Monte-Carlo simulations to ensure convergence.

We point out, however, that our method can remain a useful indicator with a limited number of Monte-Carlo simulations. This is because that a highly accurate estimate of influence power is not always necessary to solve the AUF problem; in many cases, a rough approximation is sufficient.

For instance, if a variable's true influence is non-positive ($\dot{p} \leq 0$), the approximation succeeds as long as it correctly suggests that no alteration is beneficial. Formally, this only requires the approximated MEP terms to satisfy $\hat{p}_{N_0,\emptyset} \geq \max_{c \in \Delta_{N_0}} \hat{p}_{N_0,c}$. Similarly, if the true influence is positive ($\dot{p} > 0$), the approximation succeeds as long as it correctly identifies the optimal alteration c^*, which is possible if relative benefits are roughly gauged. Formally, this only requires the approximated MEP terms to satisfy $\hat{p}_{N_0,c^*} \geq \max_{c \in \Delta_{N_0}} \hat{p}_{N_0,c}$. Therefore, even if the approximation is not perfect, our method could still provide valuable guidance, helping practitioners to prioritize alterations and focus on the variables most likely to be influential under limitied computational resources.

5 EXPERIMENTS

In this section, we conduct experiments to validate the effectiveness of influence power. The results show that the influence power is informative for determining whether and how to alter variables without relying heavily on precise estimates of probability quantities, and the proposed method leveraging influence power outperforms existing methods on AUF tasks.

We simulate three tasks including TRADER, FARMER, and DOCTOR. For each task, we generate 1000 samples from the underlying SCM to form the observational data and repeat experiments with 10 times. The details of the tasks are provided in Appendix A due to space limitation. In each task, we consider three different methods for selecting alterations: (1) MAX-ONE: selecting the single variable with the highest AUF probability for alteration, as described in Equation (1); (2) MAX-ALL: selecting all alterable variables for alteration, as described in Equation (2); and (3) OURS: using MCTS to search for determining whether and how to alter variables based on influence power, with different numbers of iterations T. The parameter α is set to $\sqrt{2}$ by following Kocsis & Szepesvári (2006). For fair comparison, the feasible domain of alteration for each alterable variable to be $\{0, 1\}$ and the number of alterable variables is set to 3 for all methods. The performance of each method is evaluated by probability of success, i.e., the probability that the target variable falls into the desired region after performing alterations on the suggested variables.

Figure 2 shows the convergence of approximation of influence power using MCTS. We plot the deviation of the approximated influence power for the first alterable variable in each task, which is the

absolute difference with the exact value of influence power. The sign error rate indicates the frequency of inconsistency between the sign of the approximated value and the exact value. The overall error rate also takes into account of the optimality of the alteration value when the sign of approximated influence power is positive. We observe that in all cases, both the deviation and error rate decrease as T increases, demonstrating the effectiveness of MCTS in approximating influence power. Notably, the deviation continues to decrease after the error rate has converged to zero, demosntrating that our method could be effective when the MEP terms are not approximated very perfectly.

Table 1 compares our method with baselines. We observe that our method consistently outperforms both MAX-ONE and MAX-ALL across all tasks. For instance, in the FARMER task, when $T = 10$, our method achieves a probability of success of 71.57%, which is 11.88 percentage points higher than MAX-ALL and 22.66 percentage points higher than MAX-ONE. As T increases, the performance of our method improves further, reaching 74.22% when $T = 250$. Similar trends are observed in the other two tasks, with our method consistently achieving higher probability of success. These results demonstrate the superiority of the proposed method in guiding alterations for AUF tasks.

6 RELATED WORK

The rehearsal paradigm was introduced by Zhou (2022), building on the concept of influence (Zhou, 2023), This paradigm advocates for mentally simulating future possibilities in order to find alterations that positively influence the future target before making a final decision. This is analogous to how human cognitive process prepares for future events (Driskell et al., 1994). Motivated by this, Qin et al. (2023) proposed the first rehearsal learning approach, wherein the restriction of directionality is relaxed and *structural rehearsal models* capable of accommodating bi-directional interactions are developed. Several subsequent studies have addressed issues such as non-stationarity and non-linearity in rehearsal learning (Du et al., 2024; Qin et al., 2025), requiring that the structure of the underlying equations are provided by experts. Besides, while the forward-looking decision-making setting is also conceptually related to markov decision processes in reinforcement learning (Sutton & Barto, 2018), a key distinction is that the AUF problem operates under a "no going back" constraint. Unlike in many RL settings where an agent can revisit states, the past variables cannot be changed in our context. Our approximation method is particularly inspired by Monte Carlo Tree Search (MCTS) (Browne et al., 2012), which excel at planning in large state spaces by simulating future trajectories, making them well-suited for the challenges of the AUF problem.

Many efforts have been dedicated to identify causal structures and causal effects from observational data in the literature (Verma & Pearl, 1991; Cooper & Herskovits, 1992; Heckerman et al., 1995; Zheng et al., 2018; Lorch et al., 2021). We also note that researchers have proposed various ways of quantifying the strength of causal contributions, sometimes referred to as "causal influence" (Rosenbaum & Rubin, 1983; Holland, 1988; Janzing et al., 2013; Heskes et al., 2020). Different notions of influence coexist for good reason, as they formalize different perspectives on different goals (Janzing et al., 2024). Much of the prior work has focused on quantifying intrinsic causal contributions—that is, the degree to which various factors "explain" the variance of a target variable. This is valuable for attribution and scientific understanding. Our work, in contrast, focuses on quantifying practical utility for decision-making in the AUF problem. This paper primarily focuses on comparing with average causal effects, a most commonly used measure of causal strength; the comparison regarding other measures of causal strength would be similar and left for future work.

7 CONCLUSION

In this paper, we attempt to quantify the influence of alterable variables in avoiding undesired future. Drawing on intuitive considerations within the AUF framework, we introduce a novel quantity called influence power, designed to assess the extent to which variables can be manipulated to increase the AUF probability. Our investigation uncovers an intriguing possibility that non-causal variables can have non-trivial influence power on the future target, and shows that the influence power remains meaningful with imprecise approximation of probabilities. We pinpoint the obstacles in exactly evaluating the proposed measure and provide a Monte Carlo-based method for efficiently approximating it using observational data. Experiments on simulated tasks validate effectiveness of influence power for suggesting alterations to address the AUF problem.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 1957.

Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25(147):1–7, 2024. URL `http://jmlr.org/papers/v25/22-1258.html`.

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9:309–347, 1992.

James E Driskell, Carolyn Copper, and Aidan Moran. Does mental practice enhance performance? *Journal of applied psychology*, 79(4):481, 1994.

Wen-Bo Du, Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Avoiding undesired future with minimal cost in non-stationary environments. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Wen-Bo Du, Hao-Yi Lei Lei, Lue Tao, Tian-Zuo Wang, and Zhi-Hua Zhou. Enabling optimal decisions in rehearsal learning under care condition. In *International Conference on Machine Learning*, 2025.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197–243, 1995.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Advances in neural information processing systems*, volume 33, pp. 4778–4789, 2020.

Paul W Holland. Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1):i–50, 1988.

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

Dominik Janzing, Patrick Blöbaum, Atalanti A Mastakouri, Philipp M Faller, Lenon Minorics, and Kailash Budhathoki. Quantifying intrinsic causal contributions via structure preserving interventions. In *International Conference on Artificial Intelligence and Statistics*, pp. 2188–2196. PMLR, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293. Springer, 2006.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. *Advances in Neural Information Processing Systems*, 34:24111–24123, 2021.

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

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

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

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

Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Gradient-based nonlinear rehearsal learning with multivariate alterations. In *AAAI conference on artificial intelligence*, volume 39, 2025.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 4 edition, 2020.

Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence*, pp. 255–270, 1991.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

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, 2023.

A DETAILED SETTINGS

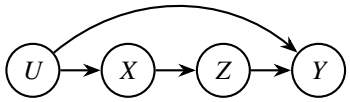
Our experiments are conducted using Intel Xeon E-2288G CPUs, featuring 8 cores and 16 threads with a frequency of 3.7 GHz. The implementation is based on DOWHY (Sharma & Kiciman, 2020; Blöbaum et al., 2024). The code to reproduce our results will be made publicly available.

A.1 THE TRADER TASK

The underlying structural equations in the TRADER Task are described in Example 5. For a trader, both the economic climate (U) and marketing strategy (Z) are important for final quarterly profit (Y). During an economic recession ($U = 1$), the initial consumer demand (X) is naturally lower, which affects the choice of marketing strategy (Z) through intermediate adjustments.

Example 5. *Consider the following structural equations with the corresponding causal graph:*

$$\begin{aligned} U &:= N_U, \\ X &:= U \cdot N_X + (1-U) \cdot (1-N_X), \\ Z &:= X \cdot N_Z + (1-X) \cdot (1-N_Z), \\ Y &:= Z \cdot (1-U) + (1-Z) \cdot N_Y, \end{aligned}$$



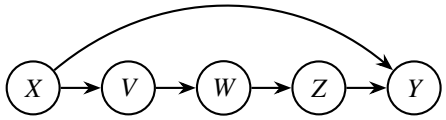
where $N_U, \sim \text{Bern}(0.5)$*,* $N_X, N_Z \overset{iid}{\sim} \text{Bern}(0.9)$*, and* $N_Y \sim \text{Bern}(0.4)$*. Let* X *and* Z *be alterable variables with* $\Delta_X = \{0, 1\}$ *and* $\Delta_Z = \{0, 1\}$*, and let the desired region be* $\mathcal{S} = \{1\}$*.*

A.2 THE FARMER TASK

The underlying structural equations in the FARMER Task are described in Example 6. For a farmer, both light (Z) and water (Z) are both important for crop yields (Y). When sunlight is abundant, it will affect the situation of water Z through intermediate variables (V and W).

Example 6. *Consider the following structural equations with the corresponding causal graph:*

$$\begin{aligned} X &= N_X, \\ V &= (1-X) \cdot N_V, \\ W &= (1-V) \cdot N_W, \\ Z &= (1-W) \cdot N_Z, \\ Y &= X \cdot Z \cdot N_Y, \end{aligned}$$



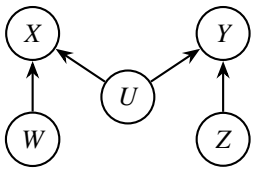
where $N_X, N_V, N_W, N_Z, N_Y \overset{iid}{\sim} \text{Bern}(0.9)$*. Alterable variables include* V*,* W*, and* Z*, whose alterable domains are all specified as* $\{0, 1\}$*. The desired region for* Y *is* $\mathcal{S} = \{1\}$*.*

A.3 THE DOCTOR TASK

The underlying structural equations in the DOCTOR Task are described in Example 7. Suppose that a doctor diagnosed a patient with seasonal flu, and the doctor had developed a fast-acting drug (Z). W denotes the administration of a skin test, and X denotes the skin response. Together, W and X would reflects the information of U, the allergy gene. The target variable Y denotes the state of recovery.

Example 7. *Consider the following structural equations with the corresponding causal graph:*

$$\begin{aligned} U &= N_U, \\ W &= N_W, \\ X &= U \cdot W \cdot (1-N_X), \\ Z &= N_Z, \\ Y &= Z \cdot (1-U) + (1-Z) \cdot N_Y, \end{aligned}$$



where $N_U \sim \text{Bern}(0.5)$*,* $N_W, N_X, N_Z \overset{iid}{\sim} \text{Bern}(0.1)$*, and* $N_Y \sim \text{Bern}(0.4)$*. Let* W*,* X*, and* Z *be alterable variables with* $\Delta_W = \Delta_X = \Delta_Z = \{0, 1\}$*, and let the desired region be* $\mathcal{S} = \{1\}$*.*

B PROOF OF PROPOSITION 1

Recall from Equation (9), the joint distribution conditioned on the alteration set $\hat{\mathbf{A}}$ is expressed as $P(\mathbf{X}|\hat{\mathbf{A}}) = \prod_{X_i \in \mathbf{A}} \delta(X_i) \prod_{X_i \in \mathbf{X} \setminus \mathbf{A}} P(X_i|X_1, \ldots, X_{i-1})$. Because the sequence is topologically consistent with the underlying SCM, and the SCM is assumed to be acyclic, the value of each variable X_i depends solely on its direct parents PA_i; consequently, $P(X_i|X_1, \ldots, X_{i-1}) = P(X_i|\mathrm{PA}_i)$. Substituting this observation back into the product shows that $P(\mathbf{X}|\hat{\mathbf{A}}) = \prod_{X_i \in \mathbf{A}} \delta(X_i) \prod_{X_i \in \mathbf{X} \setminus \mathbf{A}} P(X_i|\mathrm{PA}_i)$. By invoking the manipulation theorem (i.e., Theorem 3.6 in Spirtes et al. (2000)), we conclude that Equation (9) is precisely the probability of $\mathbf{X}$ under alteration of $\mathbf{A}$. Moreover, the quantity $P(Y \in \mathcal{S}|\hat{\mathbf{A}}, \mathbf{O})$ in Equation (10) is fully determined by $P(\mathbf{X}|\hat{\mathbf{A}})$, and therefore Equation (10) indeed gives to the true AUF probability dictated by the underlying SCM.