

Agent-Specific Effects

Stelios Triantafyllou¹, Aleksa Sukovic^{1,2}, Debmalya Mandal¹, Goran Radanovic¹

¹Max Planck Institute for Software Systems, ²Saarland University
 {strianta, asukovic, dmandal, gradanovic}@mpi-sws.org

Abstract

Establishing causal relationships between actions and outcomes is fundamental for accountable multi-agent decision-making. However, interpreting and quantifying agents' contributions to such relationships pose significant challenges. These challenges are particularly prominent in the context of multi-agent sequential decision-making, where the causal effect of an agent's action on the outcome depends on how the other agents respond to that action. In this paper, our objective is to present a systematic approach for attributing the causal effects of agents' actions to the influence they exert on other agents. Focusing on multi-agent Markov decision processes, we introduce agent-specific effects (ASE), a novel causal quantity that measures the effect of an agent's action on the outcome that propagates through other agents. We then turn to the counterfactual counterpart of ASE (cf-ASE), provide a sufficient set of conditions for identifying cf-ASE, and propose a practical sampling-based algorithm for estimating it. Finally, we experimentally evaluate the utility of cf-ASE through a simulation-based testbed, which includes a sepsis management environment.

1 Introduction

Conducting post-hoc analysis on decisions and identifying their causal contributions to the realized outcome represents a cornerstone of accountable decision making. Such analysis can be particularly important in safety-critical applications, where accountability is paramount. Often, these applications involve multiple decision makers or agents, who interact over a period of time, and make decisions or actions that are interdependent. This, in turn, implies that the impact of an agent's action on the outcome depends on how other agents respond to that action in subsequent periods. The interdependence between agents' actions poses challenges in quantifying and interpreting their causal contributions, since their effects are interleaved rather than isolated.

To illustrate these challenges, consider a canonical problem setting where accountability is paramount: the sepsis treatment problem (Komorowski et al. 2018). In a multi-agent variant of this problem, a human clinician and an AI agent decide on a series of treatments for an ICU patient over a period of time. For example, the AI can function as the primary decision maker, suggesting treatments that can

be overridden by the clinician, who assumes a supervisory role. A simplified, one-shot decision making instance of this problem is depicted in Fig. 1 (left) through a causal graph.

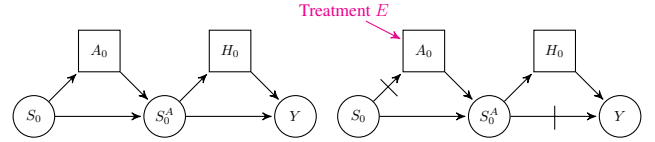


Figure 1: The figure illustrates the sepsis treatment problem setting over 1 time-step. Squares denote agents' actions, A for AI and H for clinician. Circles S are patient states, while S^A are the conjunctions of S and A . Y is the outcome. Edges that are struck through represent deactivated edges. Exogenous (magenta) arrows represent interventions on A_0 that fix its value to the action indicated. The graph on the right depicts a path-specific effect of treatment E on Y .

Given the causal graph, we can utilize standard causal concepts to quantify the impact of the agents' actions on the outcome Y . Arguably, the most straightforward way of doing so would be through the notion of *total counterfactual effects*. Given, for example a scenario where the AI suggests a treatment C and an undesirable outcome is realized, the total counterfactual effect can be used to measure the impact that an alternative treatment E would have had on the outcome. Evaluating this effect involves intervening on the AI's action, setting it to treatment E , and measuring the probability that the undesirable outcome would have been avoided. However, this analysis may be misleading for assessing the accountability of the AI agent, as it is unclear to which degree its action affects the outcome through the clinician.

To quantify the effect of the AI's action that propagates through the clinician, we can utilize the concept of *path-specific effects* (Pearl 2001). In particular, we can measure the total counterfactual effect of treatment E on outcome Y in a modified model where the edge between S_0^A and Y in Fig. 1 (right) is deactivated. The latter can be interpreted as if the outcome will be realized based on the factual action of the AI, i.e., treatment C , and the action taken by the clinician, who sees the alternative action, i.e., treatment E .

A natural generalization of this setting to multiple time-steps is depicted in Fig. 2. To ensure that action A_1 of the

AI is the same as the factual one, and hence that the AI’s response in the next time-step does not affect the counterfactual outcome, we could also deactivate the edge between S_1 and A_1 . While intuitively appealing, the effect measured in this newly modified model does not capture higher-order dependencies between the agents’ actions. Namely, the clinician’s policy depends on the AI’s policy, so by fixing the AI’s subsequent actions to the ones originally taken, we inadvertently remove such a dependency from the analysis.

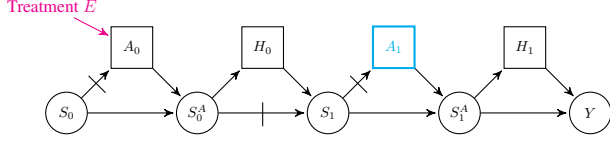


Figure 2: The causal graph depicts a path-specific effect of treatment E on Y in the sepsis treatment setting over 2 time-steps. It aims to capture the counterfactual effect of E that propagates through the clinician. A cyan colored node signifies that the node is set to the action that the agent took in the factual scenario, i.e., under treatment C .

Our objective is to capture these dependencies through a novel causal notion termed *agent-specific effects*. At a high level, we contend that the clinician should behave as if the AI’s actions are not fixed, but rather responsive to the clinician’s decisions. However, the effect should be measured using the factual actions of the AI. We show that analyzing this effect necessitates counterfactual reasoning across three distinct scenarios, and hence it cannot be expressed through path-specific effects. Our contributions are as follows.

Agent-Specific Effects. (Section 3) Focusing on multi-agent Markov decision processes (MMDPs), we introduce (*counterfactual*) *agent-specific effects*, a novel causal quantity that measures the effect of an agent’s action on the MMDP outcome that propagates through a selected subset of agents. To enable counterfactual reasoning we represent MMDPs by structural causal models (Section 2).

Identifiability. (Section 4) We study the identifiability of agent-specific effects. Even though we show that they are in general non-identifiable, we overcome this issue by considering an existing causal property, which we refer to as *noise monotonicity*. We show that assuming noise monotonicity suffices for the identification of counterfactual agent-specific effects. Furthermore, recognizing the scarcity of theoretical analysis for noise monotonicity, we present novel findings to enhance its characterization.

Algorithm. (Section 5) We propose a practical sampling-based algorithm for the estimation of counterfactual agent-specific effects. We show that this algorithm outputs unbiased estimates provided that noise monotonicity holds.

Connections to PSE. (Section 3 and 6) We discuss in detail the importance of agent-specific effects and their conceptual benefits relative to path-specific effects. We introduce a generalized notion of path-specific effects, which allows us to establish a formal connection between agent-specific effects and path-specific effects.

Experiments. (Section 7) We conduct extensive experi-

ments on two test-beds, *Graph* and *Sepsis*, evaluating the robustness and practicality of our approach.¹

1.1 Additional Related Work

This paper is related to works on *mediation* or *path analysis* with a single (Pearl 2001; Zhang and Bareinboim 2018b,a) or multiple (Daniel et al. 2015; Steen et al. 2017; Avin, Shpitser, and Pearl 2005; Singal, Michailidis, and Ng 2021; Chiappa 2019) mediators. Unlike this line of work, we do not study the effect of an exposure variable (in our case action) on a response variable (in our case outcome) that propagates through a specific variable or path in a general causal model. Instead, we focus on a specific framework, MMDPs, and we aim to quantify and reason about effects that propagate through a subset of decision-making agents, who control multiple sequentially dependent actions.

Our work also relates to the area of *causal contributions*, which studies the problem of attributing to multiple causes a contribution to a target effect (Jung et al. 2022), e.g., to explain the effect of different model features on predictions (Heskes et al. 2020). Even though we do not develop here such attribution methods, we view this as a natural next step.

Close to our work are also papers that look into the problem of measuring the *influence* between actions in multi-agent decision making systems (Jaques et al. 2019; Pieroth, Fitch, and Belzner 2022). The main distinction between this type of works and ours is that we care about the influence of an action on the outcome through the actions of other agents, rather than merely the influence of one action on another.

From a technical point of view, this paper is also close to works which combine SCMs with sequential decision making frameworks similar to MMDPs, e.g., MDPs (Tsirtsis and Gomez-Rodriguez 2023), POMDPs (Buesing et al. 2018) and Dec-POMDPs (Triantafyllou and Radanovic 2023).

Finally, this paper is related to works that utilize causal tools to study accountability or related concepts in decision making. We consider works on *responsibility* in multi-agent one-shot (Chockler and Halpern 2004) or sequential (Triantafyllou, Singla, and Radanovic 2022) decision making, as well as works that focus on *explainability* (Madumal et al. 2020), *incentives* (Everitt et al. 2021; Farquhar, Carey, and Everitt 2022), *harm* (Richens, Beard, and Thompson 2022), (Carroll et al. 2023) and *intention* (Halpern and Kleiman-Weiner 2018). Our view is that agent-specific effects can be integrated into all aforementioned works enhancing their (causal) reasoning, e.g., by modifying current responsibility attribution methods to account for the effect of an *actual cause* that propagates through other agents, or by defining agent-specific counterfactual harm to measure the harm caused by an action only through a set of agents.

2 Framework and Preliminaries

In this section, we introduce and establish a connection between the two main parts of our formal framework, Structural Causal Models (SCMs) (Pearl 2009) and Multi-Agent Markov Decision Processes (MMDPs) (Boutilier 1996).

¹Code to reproduce the experiments is available at <https://github.com/stelios30/agent-specific-effects.git>.

2.1 Graphs

We begin by introducing some necessary graph notation. For a directed graph G , we denote with $\mathbf{V}(G)$ and $\mathbf{E}(G)$ the sets of vertices and edges in G . With $\mathbf{E}_+^{V^i}(G)$ or $\mathbf{E}_+^i(G)$ for short, we denote the set of outgoing edges of node V^i in G . Furthermore, we use $Pa^i(G)$ to denote the parent set of V^i in G . We call graph g an *edge-subgraph* of G if $\mathbf{V}(g) = \mathbf{V}(G)$ and $\mathbf{E}(g) \subseteq \mathbf{E}(G)$. We also denote with \bar{g} the edge-subgraph of G with $\mathbf{E}(\bar{g}) = \mathbf{E}(G) \setminus \mathbf{E}(g)$.

2.2 Structural Causal Models

An SCM M is a tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$, where: \mathbf{U} is a set of unobservable noise variables with joint probability distribution $P(\mathbf{u})$; \mathbf{V} is a set of observable random variables; \mathcal{F} is a collection of functions. Each random variable $V^i \in \mathbf{V}$ corresponds to exactly one noise variable $U^i \in \mathbf{U}$ and one function $f^i \in \mathcal{F}$. The value of V^i is determined by the *structural equation* $V^i := f^i(\mathbf{PA}^i, U^i)$, where $\mathbf{PA}^i \subseteq \mathbf{V} \setminus V^i$. Moreover, SCM M induces a *causal graph* G , where \mathbf{V} constitutes the set of nodes $\mathbf{V}(G)$, and for every node $V^i \in \mathbf{V}(G)$ set \mathbf{PA}^i constitutes its parent set $Pa^i(G)$.

Interventions. An *intervention* on variable $X \in \mathbf{V}$ in SCM M describes the process of modifying the structural equation which determines the value of X in M . We consider two types of interventions, *hard interventions* and *natural interventions*. A **hard intervention** $do(X := x)$ fixes the value of X in M to a constant $x \in \text{dom}\{X\}$.² The new SCM that is derived from performing $do(X := x)$ in M is denoted by $M^{do(X:=x)}$, and is identical to M except that function f^X is replaced by the constant function.

Now, let $Z \in \mathbf{V}$. We denote with $Z_{do(X:=x)}$ or Z_x for short, the random variable which, for some noise $\mathbf{u} \sim P(\mathbf{u})$, takes the value that Z would take in $M^{do(X:=x)}$. A **natural intervention** $do(Z := Z_x)$ replaces the function f^Z of M with Z_x . The derived SCM is denoted by $M^{do(Z:=Z_x)}$.

Note that interventions on a set of variables $\mathbf{W} \subseteq \mathbf{V}$ instead of a singleton are defined in a similar way.

Distributions. Joint distribution $P(\mathbf{V})$ is referred to as the *observational distribution* of M . Furthermore, distributions of the form $P(Y_x)_M$ and of the form $P(Y_x|\mathbf{v})_M$, where $\mathbf{v} \in \text{dom}\{\mathbf{V}\}$ and $\mathbf{v}(X) \neq x$,³ are referred to as *interventional* and *counterfactual distributions*, respectively. Subscript M here implies that the probability distribution is defined over SCM M . Note that we may drop this subscript when the underlying SCM is assumed or obvious.

Identifiability. Identifying whether or not an interventional or counterfactual distribution over SCM M can be uniquely estimated from the causal graph and observational distribution of M is a fundamental challenge in causality. Formally, a distribution $P(\cdot)_M$ is *identifiable* if $P(\cdot)_{M_1} = P(\cdot)_{M_2}$ for any two SCMs M_1 and M_2 , s.t. (a) the causal graphs induced by M_1 and M_2 are identical to that of M , and (b) the observational distributions of M_1 and M_2 agree with M 's. Otherwise, we say that $P(\cdot)_M$ is *non-identifiable*.

² $\text{dom}\{X\}$ here denotes the domain of variable X .

³ $\mathbf{v}(X)$ here denotes the value of variable X in \mathbf{v} .

Effects. The *total causal effect* of intervention $do(X := x)$ on $Y = y$, relative to reference value x^* , is defined as

$$\text{TCE}_{x,x^*}(y)_M = P(y_x)_M - P(y_{x^*})_M,$$

where $P(y_x)_M$ is short for $P(Y_x = y)_M$.

Given a realization \mathbf{v} of \mathbf{V} ,⁴ the *total counterfactual effect* of intervention $do(X := x)$ on $Y = y$ is defined as

$$\text{TCFE}_x(y|\mathbf{v})_M = P(y_x|\mathbf{v})_M - P(y|\mathbf{v})_M.$$

Assumptions. Throughout this paper, we assume SCMs to be *recursive*, i.e., their causal graphs are acyclic. We further make the *exogeneity* assumption: noise variables are mutually independent, $P(\mathbf{u}) = \prod_{u^i \in \mathbf{u}} P(u^i)$. We also assume that the domains of observable variables are finite and discrete, and that noise variables take numerical values.

2.3 Multi-Agent Markov Decision Processes

An MMDP is a tuple $\langle \mathcal{S}, \{1, \dots, n\}, \mathcal{A}, T, h, \sigma \rangle$, where: \mathcal{S} is the state space; $\{1, \dots, n\}$ is the set of agents; $\mathcal{A} = \times_{i=1}^n \mathcal{A}_i$ is the joint action space, with \mathcal{A}_i being the action space of agent i ; $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the transitions probability function; h is the finite time horizon; σ is the initial state distribution.⁵ For example, if the agents take joint action \mathbf{a} in state s , then the environment transitions to state s' with probability equal to $T(s'|s, \mathbf{a})$. We denote with π_i the policy of agent i and with π the agents' joint policy. As such, the probability of agent i taking action a_i in state s is given by $\pi_i(a_i|s)$. Furthermore, we call trajectory a realization τ of the MMDP, i.e., τ consists of a sequence of state-joint action pairs $\{(s_t, \mathbf{a}_t)\}_{t \in \{0, \dots, h-1\}}$ and a final state s_h .

Next, we describe how to model an MMDP coupled with a joint policy π as an SCM M , henceforth called MMDP-SCM. First, we define the set of observable variables \mathbf{V} to include all state and action variables of the MMDP. To each such variable, we assign a noise variable U and a function f . We then formulate the structural equations of M as

$$\begin{aligned} S_0 &= f^{S_0}(U^{S_0}); \quad S_t = f^{S_t}(S_{t-1}, \mathbf{A}_{t-1}, U^{S_t}); \\ A_{i,t} &= f^{A_{i,t}}(S^t, U^{A_{i,t}}). \end{aligned} \quad (1)$$

The causal graph of M is included in the Supplemental Material. Note that (Buesing et al. 2018) shows that such parameterization of a partially observable MDP to an SCM is always possible. Their result can be directly extended to MMDPs. This means that there always exists an MMDP-SCM M , which induces the agents' policies, and the initial state distribution and transition dynamics of the MMDP.

3 Formalizing Agent-Specific Effects

In this section, we introduce the notion of *agent-specific effects* (ASE) and its counterfactual counterpart (cf-ASE). For a given MMDP-SCM M , the purpose of ASE and cf-ASE is to quantify the total causal and counterfactual effect, respectively, of an action $A_{i,t} = a_{i,t}$ on a response variable Y , only through a selected set of agents, also called the *effect*

⁴The reference value x^* here is considered to be $\mathbf{v}(X)$.

⁵For ease of notation, rewards are considered part of the state.

agents. To that end, our definitions consider **three alternative worlds**. In the first world, action $a_{i,t}$ is fixed. In the second world, a reference $a_{i,t}^*$ (for ASE) or realized $\tau(A_{i,t})$ (for cf-ASE) action is fixed instead. In the third world, the reference or realized action is fixed, and additionally for every time-step $t' > t$ the effect agents are forced to take the actions that they would naturally take in the first world, while all other agents, also called the *non-effect agents*, are forced to take the actions that they would naturally take in the second world. We consider response variable Y to be any state variable in M that chronologically succeeds action $A_{i,t}$.

We formally define ASE and cf-ASE based on the language of natural interventions as follows.⁶

Definition 1 (ASE). *Given an MMDP-SCM M and a non-empty subset of agents \mathbf{N} in M , the \mathbf{N} -specific effect of intervention $do(A_{i,t} := a_{i,t})$ on $Y = y$, relative to reference action $a_{i,t}^*$, is defined as*

$$\text{ASE}_{a_{i,t}, a_{i,t}^*}^{\mathbf{N}}(y)_M = P(y_{a_{i,t}^*})_{M^{do(I)}} - P(y_{a_{i,t}^*})_M,$$

where $I = \{Z := Z_{a_{i,t}^*} | Z \in \{A_{i',t'}\}_{i' \notin \mathbf{N}, t' > t}\} \cup \{Z := Z_{a_{i,t}} | Z \in \{A_{i',t'}\}_{i' \in \mathbf{N}, t' > t}\}$.

Definition 2 (cf-ASE). *Given an MMDP-SCM M , a non-empty subset of agents \mathbf{N} in M and a trajectory τ of M , the counterfactual \mathbf{N} -specific effect of intervention $do(A_{i,t} := a_{i,t})$ on $Y = y$ is defined as*

$$\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M = P(y_{\tau(A_{i,t})}|\tau; M)_{M^{do(I)}} - P(y|\tau)_M,$$

where $I = \{A_{i',t'} := \tau(A_{i',t'})\}_{i' \notin \mathbf{N}, t' > t} \cup \{Z := Z_{a_{i,t}} | Z \in \{A_{i',t'}\}_{i' \in \mathbf{N}, t' > t}\}$.

Going back to the sepsis problem from Section 1, the clinician here takes the role of the effect agent ($i' \in \mathbf{N}$) and the AI that of the non-effect agent ($i' \notin \mathbf{N}$). The (counterfactual) clinician-specific effect of intervention $do(A_0 := E)$ on Y is illustrated in Fig. 3. We see that in this model, the AI takes the actions that it would naturally take under the intervention $do(A_0 := C)$, i.e., its original actions, while the clinician takes the actions that it would naturally take under the intervention $do(A_0 := E)$. This should be contrasted with the path-specific effect in Fig. 2, where the clinician behaves as if the AI agent takes its original action in A_1 .

To demonstrate the importance of this distinction, consider a clinician who does not account for AI's actions when deciding whether or not to override them, but instead bases its decisions solely on the patient's state, e.g., assuming full control whenever the patient is in critical condition. Under this behavior, the path-specific effect in Fig. 2 always evaluates to 0. Namely, the intervention on A_0 does not impact the realization of S_1 : the clinician does not respond to the AI's action, while the edge between S_0^A and S_1 is deactivated. Hence, the realized states and actions in later periods are the same as in the factual world. We deem this to be problematic since we block the influence that the intervention on A_0

may have on the clinician's actions in later periods (e.g., H_1 in Fig. 2) through state S_1 . On the other hand, for the agent-specific effect in Fig. 3, the clinician acts according to the states that naturally realize under intervention $do(A_0 := E)$, so this influence is not blocked. For example, if the initial state is such that the clinician does not decide to override, the realization of state S_1 under intervention $do(A_0 := E)$ could be possibly different than the one in the factual world.

While it suffices to reactivate the edge between S_0^A and S_1 in Fig. 2 to obtain a path-specific effect that alleviates the aforementioned issue, this modified version can lead to a different type of counter-intuitive results. Namely, in a scenario where the clinician never overrides the AI's actions, this measure might have a positive evaluation. Our experiments in Section 7 demonstrate this tendency. In contrast, the agent-specific effect in Fig. 3 does not suffer from this issue since we intervene on A_0 with $do(A_0 := C)$.

Despite these distinctions, we can formalize both of these quantities using a generalized notion of path-specific effects, which we introduce in Section 6.

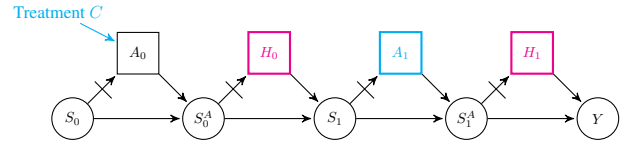


Figure 3: The causal graph depicts the agent-specific effect of treatment E on Y in the sepsis problem from the introduction. A magenta (resp. cyan) colored node signifies that the node is set to the action that the agent would naturally take under intervention E (resp. the action that the agent took in the factual scenario, i.e., under treatment C).

4 Identifiability Results

Note that $P(y|\tau)_M$ from Definition 2 is 1 if $y = \tau(Y)$, and 0 otherwise. We conclude then that $P(y|\tau)_M$ is identifiable. In contrast, $P(y_{\tau(A_{i,t})}|\tau; M)_{M^{do(I)}}$ is non-identifiable, since evidence $\mathbf{V} = \tau$ is contradictory to intervention $do(I)$ (Shpitser and Pearl 2008). Subsequently, $\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M$ is also non-identifiable under our current set of assumptions.

Theorem 1 states that $\text{ASE}_{a_{i,t}, a_{i,t}^*}^{\mathbf{N}}(y)_M$, which is defined over interventional probabilities instead of counterfactual ones, is also non-identifiable.⁷ The proof is based on the *re-acting witness criterion* (Avin, Shpitser, and Pearl 2005).

Theorem 1. *Let M be an MMDP-SCM, and \mathbf{N} be a non-empty subset of agents in M . It holds that the \mathbf{N} -specific effect $\text{ASE}_{a_{i,t}, a_{i,t}^*}^{\mathbf{N}}(y)_M$ is non-identifiable if $t_Y > t + 1$, where t_Y denotes the time-step of variable Y .*

To overcome the aforementioned non-identifiability issues, we consider the following property, which was first introduced in (Pearl 2009, p. 295) for binary SCMs.

⁶We note that our main theoretical and experimental results are w.r.t. cf-ASE. However, we also provide the definition of ASE for completeness and to enable a direct comparison with existing causal quantities (Section 6).

⁷Note that the same non-identifiability results would also hold even if we had access to interventional data.

Definition 3 (Noise Monotonicity). *Given an SCM M with causal graph G , we say that variable $V^i \in \mathbf{V}$ is noise-monotonic in M w.r.t. a total ordering \leq_i on $\text{dom}\{V^i\}$, if for any $pa^i \in \text{dom}\{Pa^i(G)\}$ and $u_1^i, u_2^i \sim P(U^i)$ s.t. $u_1^i < u_2^i$, it holds that*

$$f^i(pa^i, u_1^i) \leq_i f^i(pa^i, u_2^i).$$

Assuming *noise monotonicity* means that functions f^i of the underlying SCM, or MMDP-SCM in our case, are monotonic w.r.t. noise variables U^i . The next theorem states that noise monotonicity suffices to render cf-ASE identifiable.⁸

Theorem 2. *Let M be an MMDP-SCM, \mathbf{N} be a non-empty subset of agents in M , and τ be a trajectory of M . The counterfactual \mathbf{N} -specific effect $\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M$ is identifiable, if for every variable $Z \in \mathbf{V}$ there is a total ordering \leq_Z on $\text{dom}\{Z\}$ w.r.t. which Z is noise-monotonic in M .*

Theorem 2 is similar to Theorem 1 from (Lu et al. 2020), and is stated here for completeness. Our proof is based on a lemma that allows to express counterfactual probabilities as functions of the model’s observational distribution.

Despite the importance of noise monotonicity, enabling the identification of counterfactual probabilities, it remains devoid of comprehensive theoretical analysis. Thus, for the remainder of this section, we present novel findings to enhance its characterization. (Buesing et al. 2018) show how we can represent any partially observable MDP by an SCM. In that direction, we show next how to represent any MMDP by an MMDP-SCM that satisfies noise monotonicity.

Lemma 1. *Let $\mathbf{X} = \{X^1, \dots, X^N\}$ represent the set of state and action variables of an MMDP, and $P(\mathbf{X})$ be the joint distribution induced by the MMDP under a joint policy π . Additionally, let \leq_1, \dots, \leq_N be total orderings on $\text{dom}\{X^1\}, \dots, \text{dom}\{X^N\}$, respectively. We can construct an MMDP-SCM $M = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$ with causal graph G , where $\mathbf{V} = \{V^1, \dots, V^N\}$, $\mathbf{U} = \{U^1, \dots, U^N\}$, $P(U^i) = \text{Uniform}[0, 1]$, and functions in \mathcal{F} are defined as*

$$f^i(Pa^i(G), U^i) = \inf_{v^i} \{P(X^i \leq_i v^i | Pa^i(G))_M \geq U^i\},$$

such that $P(\mathbf{X}) = P(\mathbf{V})$ and every variable $V^i \in \mathbf{V}$ is noise-monotonic in M w.r.t. \leq_i .

(Oberst and Sontag 2019) recently proposed another property for categorical SCMs, *counterfactual stability*. Interestingly, they show that their property implies the well-known property of *monotonicity* (Pearl 1999) in binary SCMs. Note, that counterfactual stability does not suffice for identifiability. Proposition 1 states that noise monotonicity also implies monotonicity in binary SCMs.

Proposition 1. *Let M be an SCM consisting of two binary variables X and Y with $Y = f^Y(X, U^Y)$. If Y is noise-monotonic in M w.r.t. the numerical ordering, then Y is monotonic relative to X in M . However, if Y is monotonic relative to X in M , then Y is not necessarily noise-monotonic in M w.r.t. any total ordering on $\text{dom}\{X\}$.*

⁸Even though we do not explicitly prove it in this paper, it follows that the same result also holds for ASE.

5 Algorithm

We now present our algorithm for the estimation of cf-ASE. Algorithm 1 follows a standard methodology for counterfactual inference (Pearl 2009): it first samples noise variables from the posterior $P(\mathbf{u}|\tau)$ (abduction), and then for each sample computes cf-ASE (prediction) by performing the interventions specified in Definition 2 (action). Finally, the algorithm outputs the average cf-ASE value overall samples.

Algorithm 1: Estimates $\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M$

Input: MMDP-SCM M , trajectory τ , effect agents \mathbf{N} , action variable $A_{i,t}$, action $a_{i,t}$, outcome variable Y , outcome y , number of posterior samples H

```

1:  $h \leftarrow 0$ 
2:  $c \leftarrow 0$ 
3: while  $h < H$  do
4:    $\mathbf{u}_h \sim P(\mathbf{u}|\tau)$  {Sample noise from the posterior}
5:    $h \leftarrow h + 1$ 
6:    $\tau^h \sim P(\mathbf{V}|\mathbf{u}_h)_{M^{do(A_{i,t}:=a_{i,t})}}$  {Sample cf trajectory}
7:    $I(\mathbf{u}_h) \leftarrow \{A_{i',t'} := \tau(A_{i',t'})\}_{i' \notin \mathbf{N}, t' > t} \cup \{A_{i',t'} := \tau^h(A_{i',t'})\}_{i' \in \mathbf{N}, t' > t}$ 
8:    $y^h \sim P(Y|\mathbf{u}_h)_{M^{do(I(\mathbf{u}_h))}}$  {Sample cf outcome}
9:   if  $y^h = y$  then
10:     $c \leftarrow c + 1$ 
11:   end if
12: end while
13: return  $\frac{c}{H} - \mathbb{1}(\tau(Y) = y)$ 

```

Theorem 3. *Let MMDP-SCMs $M = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$ and $\hat{M} = \langle \mathbf{U}, \mathbf{V}, \hat{\mathcal{F}}, \hat{P}(\mathbf{u}) \rangle$, s.t. $P(\mathbf{V})_M = P(\mathbf{V})_{\hat{M}}$ and every variable $Z \in \mathbf{V}$ is noise-monotonic in both M and \hat{M} w.r.t. the same total ordering \leq_Z on $\text{dom}\{Z\}$. It holds that the output of Algorithm 1, when it takes as input model \hat{M} , is an unbiased estimator of $\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M$.*

If noise-monotonicity holds, then according to Theorem 3, access to a model \hat{M} suffices to get an unbiased estimate of $\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M$. Based on Lemma 1, such an MMDP-SCM always exists, and it can be derived solely from the observational distribution $P(\mathbf{V})_M$ and the total orderings w.r.t. which variables in M are noise-monotonic.

6 Connection to Path-Specific Effects

Path-specific effects are a well-known causal concept used to quantify the causal effect of one variable on another along a specified set of paths in the causal graph (Pearl 2001).

Definition 4 (PSE). *Given an SCM M with causal graph G , and an edge-subgraph g of G (also called the effect subgraph), the g -specific effect of intervention $do(X := x)$ on $Y = y$, relative to reference value x^* , is defined as*

$$\text{PSE}_{x,x^*}^g(y)_M = \text{TCE}_{x,x^*}(y)_{M_g},$$

where $M_g = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_g, P(\mathbf{u}) \rangle$ is a modified SCM, which induces g and is formed as follows. Each function f^i in

\mathcal{F} is replaced with a new function f_g^i in \mathcal{F}_g , s.t. for every $pa^i(g) \in \text{dom}\{Pa^i(g)\}$ and $\mathbf{u} \sim P(\mathbf{u})$ it holds $f_g^i(pa^i(g), u^i) = f^i(pa^i(g), pa^i(\bar{g})^*, u^i)$. Here, $pa^i(\bar{g})^*$ denotes the value of $Pa^i(\bar{g})_{x^*}$ in M for noise \mathbf{u} .

As argued in Section 1, despite also considering natural interventions, path-specific effects cannot be used to express agent-specific effects. Intuitively, this is because analyzing the latter effects requires reasoning across three alternative worlds, while the definition of the former reasons across only two. To establish a deeper connection between the two concepts, we define the notion of *fixed path-specific effects* (FPSE) utilizing a similar path-deactivation process as in Definition 4.2.⁹ Supplemental Material includes identifiability results for FPSE and its counterfactual counterpart.

Definition 5 (FPSE). Given an SCM M with causal graph G , and two edge-subgraphs of G , effect subgraph g and reference subgraph g^* such that $\mathbf{E}(g) \cap \mathbf{E}(g^*) = \emptyset$, the fixed g -specific effect of intervention $do(X := x)$ on $Y = y$, relative to reference value x^* and g^* , is defined as

$$\text{FPSE}_{x,x^*}^{g,g^*}(y)_M = P(y_{x^*})_{M_q} - P(y_{x^*})_M,$$

where q is the edge-subgraph of G with $\mathbf{E}(q) = \mathbf{E}(G) \setminus \{\mathbf{E}(g), \mathbf{E}(g^*)\}$, and $M_q = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_q, P(\mathbf{u}) \rangle$ is a modified SCM, which induces q and is formed as follows. Each function f^i in \mathcal{F} is replaced with a new function f_q^i in \mathcal{F}_q , s.t. for every $pa^i(q) \in \text{dom}\{Pa^i(q)\}$ and $\mathbf{u} \sim P(\mathbf{u})$ it holds $f_q^i(pa^i(q), u^i) = f^i(pa^i(q), pa^i(g)^e, pa^i(g^*)^*, u^i)$. Here, $pa^i(g^*)^*$ (resp. $pa^i(g)^e$) denotes the value of $Pa^i(g^*)_{x^*}$ (resp. $Pa^i(g)_x$) in M for noise \mathbf{u} .

Definitions of PSE and FPSE have two main distinctions: FPSE considers one additional edge-subgraph compared to PSE, and it computes the counterfactual probability over its modified model under intervention $do(X := x^*)$ instead of $do(X := x)$. This additional subgraph allows FPSE to reason across three alternative worlds, i.e., one more than PSE, which is why it can be used to express agent-specific effects.

Lemma 2. Let M be an MMDP-SCM with causal graph G and \mathbf{N} be a non-empty subset of agents in M . It holds that

$$\text{ASE}_{a_i,t,a_{i,t}^*}^{\mathbf{N}}(y)_M = \text{FPSE}_{a_i,t,a_{i,t}^*}^{g,g^*}(y)_M$$

where g is the edge-subgraph of G with $\mathbf{E}(g) = \bigcup_{A_{i',t'}: i' \in \mathbf{N}, t' > t} \mathbf{E}_+^{A_{i',t'}}(G)$ and g^* is the edge-subgraph of G with $\mathbf{E}(g^*) = \bigcup_{A_{i',t'}: i' \notin \mathbf{N}, t' > t} \mathbf{E}_+^{A_{i',t'}}(G)$.

Furthermore, by “not using” the extra edge-subgraph of FPSE, we are able to express path-specific effects.

Proposition 2. Given an SCM M with causal graph G , and an edge-subgraph g of G , it holds that

$$\text{PSE}_{x,x^*}^g(y)_M = \text{FPSE}_{x^*,x}^{\bar{g},g^0}(y)_M + \text{TCE}_{x,x^*}(y)_M,$$

where g^0 is the edge-subgraph of G with $\mathbf{E}(g^0) = \emptyset$.

⁹We emphasize here that PSE and FPSE are defined over general SCMs, while ASE is defined explicitly over MMDP-SCMs.

7 Experiments

In this section, we empirically evaluate our approach using two environments from prior work, *Graph* (Triantafyllou, Singla, and Radanovic 2021) and *Sepsis* (Oberst and Sonntag 2019), modified to align with our objectives.

Graph. In this environment, 6 agents move in a graph with 1 initial, 3 terminal and 9 intermediate nodes split into 3 layers. All nodes of a column are connected (with directed edges) to all the next column’s nodes. Thus, agents can move on the graph going *up*, *straight* or *down*. The agents’ goal is to be equally distributed when reaching the last column. The policy of each agent $i \in \{1, \dots, 6\}$ is parameterized with a $p_i = 0.05 \cdot i$, which denotes their probability of taking a random action at any node. With $1 - p_i$ probability, agent i follows a policy that is common among all agents, and depends on the agents’ current positions on the graph. We implement this environment, such that each action variable of the underlying MMDP-SCM is noise-monotonic w.r.t. a predefined total ordering, which is *up*, *down*, *straight*.

Sepsis. In our multi-agent version of Sepsis, AI and clinician have roles similar to those described in Section 1. At each round, the AI recommends one of eight possible treatments, which is then reviewed and potentially overridden by the clinician. If the clinician does not override AI’s action, i.e., takes a *no-op* action, then the treatment selected by the AI is applied. Otherwise, a new treatment selected by the clinician is applied. The agents’ goal is to keep the patient alive for 20 rounds or achieve an earlier discharge. We learn the policies of both agents using Policy Iteration (Sutton and Barto 2018), and we train them on slightly different transition probabilities. For the clinician, this gives us their policy for selecting treatments, so we additionally need to specify their probability of overriding an action of the AI agent. Across all states, this is specified by a parameter μ , which intuitively models the level of the clinician’s trust in the AI: greater values of μ correspond to higher levels of trust.

Setup. For each environment, we generate a set of trajectories in which agents fail to reach their goal, 500 for Graph and 100 (per value of μ) for Sepsis. For each trajectory τ , we compute the TCFE of all the potential alternative actions that agents could have taken. Among these alternative actions $a_{i,t}$, we retain those that exhibit a TCFE greater than or equal to a predefined threshold θ (0.75 for Graph and 0.8 for Sepsis). Formally, we require $\text{TCFE}_{a_{i,t}}(Y = \text{success}|\tau) \geq \theta$. This process yields a total of 854 selected alternative actions for Graph and 9197 for Sepsis with a strong counterfactual effect, which serve as the basis for evaluating cf-ASE. Counterfactual effects in our experiments are computed for 100 counterfactual samples. We refer the reader to Supplemental Material for additional details.

7.1 Results

Robustness. We first assess the robustness of cf-ASE in the presence of uncertainty, using the Graph environment, for which we know the underlying causal model. Given this knowledge, we can compute the correct values of cf-ASE, which we refer to as *target values*, and we do so for all 854 alternative actions that were selected using the previously outlined process, covering all possible combinations

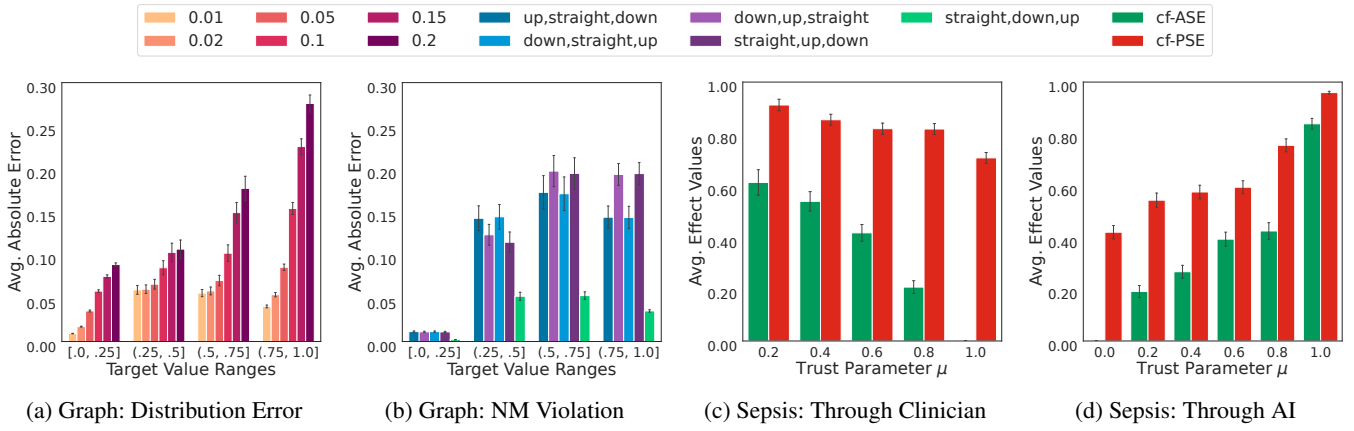


Figure 4: (Graph) Plots 4a and 4b show the average absolute error of cf-ASE when computed for wrongly estimated agents’ policies and under violation of noise monotonicity, respectively. (Sepsis) Plots 4c and 4d show average values of cf-ASE and cf-PSE when computing effects that propagate through the clinician and AI, respectively, while varying trust parameter μ .

of effect-agents (31 combinations in total). With this information, we perform two robustness tests.

First, we test the robustness to uncertainty over the observational distribution of the model, specifically the agents’ policies. The uncertainty is incorporated in the underlying model by randomly sampling a probability $\hat{p}_i \in [p_i - \epsilon_{max}, p_i + \epsilon_{max}]$, which defines i ’s (wrongly) estimated policy. Here, ϵ_{max} is a parameter that we vary. Plot 4a shows the average absolute difference between the target values and the counterfactual agent-specific effects estimated for different levels of uncertainty (i.e., ϵ_{max}). Results are grouped by target value range. We observe that for small estimation errors ($\epsilon_{max} \leq 0.05$) we can still obtain good enough estimates of cf-ASE, especially when the target value is low.

Second, we assess the impact of violating the noise monotonicity assumption on the estimation of cf-ASE, by running our estimation algorithm with an incorrect total ordering. Plot 4b shows the average absolute difference between the target values and the counterfactual agent-specific effects estimated for different misspecified total orderings. The results indicate that assuming the wrong total ordering can have a significant impact on the estimation accuracy of cf-ASE. Interestingly, when the reverse of the correct total ordering is assumed, the estimation error is relatively low.

Practicality. Next, we assess the conceptual importance and practicality of agent-specific effects using the Sepsis environment. For all selected actions of the AI (resp. clinician) we compute their counterfactual effect which propagates through the clinician (resp. AI). In Plots 4c and 4d, we present the average *clinician*- and *AI-specific effects*, as captured by cf-ASE and the counterfactual counterpart of PSE (cf-PSE), for different levels of trust. Plots 5a-5d show, for different values of trust parameter μ , the distribution of selected alternative actions across value ranges of cf-ASE.

The values shown in Plots 4c, 4d and 5a-5d suggest that the causal interdependencies between agents’ policies can have a significant impact on the total counterfactual effect of an action. For instance, in Plot 5c we can see that even for a large trust level $\mu = 0.6$, close to 40% of the AI’s

actions that admit TCFE greater than 0.8, admit a clinician-specific effect (as measured by cf-ASE) greater than 0.75. This means that the total counterfactual effect of an action can rather frequently be attributed to the behavior of other agents in the system.

Plot 4c illustrates that our approach uncovers the following intuitive pattern: the clinician-specific effect decreases as the level of trust increases, ultimately dropping to 0 when the human fully trusts the choices of the AI. This is an important finding as it implies that given a set of trajectories, cf-ASE could provide us with valuable insight about the clinician’s trust towards the AI. A similar conclusion can be drawn from Plot 4d, in which cf-ASE demonstrates the following trend: the AI-specific effect increases with the value of μ . This is expected as it implies that clinician’s actions effect less the outcome through the actions of the AI, as the latter assumes less agency. Compared to cf-ASE, we observe that cf-PSE suggests a less prominent pattern for the clinician-specific effect. Most notably, it can assign positive value to the effect in cases where the clinician blindly follows AI’s recommendations. A similar counter-intuitive result can also be observed for the AI-specific effect when $\mu = 0$. Given these results, we conclude that cf-ASE is a more suitable approach than cf-PSE for measuring the clinician- and AI-specific effects in this setting, as it aligns better with standard intuition.

Plots 5a-5d showcase the following natural trend for cf-ASE: the fraction of actions with clinician-specific effect greater than 0.75 decreases with the level of trust, while the fraction of actions with clinician-specific effect less than 0.25 increases with the level of trust. This is expected as the less the clinician trusts the AI’s actions, the more often they override them, and hence the more they tend to affect the outcome. Similarly, the fraction of actions with AI-specific effect greater than 0.75 increases with the level of trust, while the fraction of actions with AI-specific effect less than 0.25 decreases with the level of trust. This is also expected as the less the clinician trusts the AI’s actions the less agency the AI assumes.

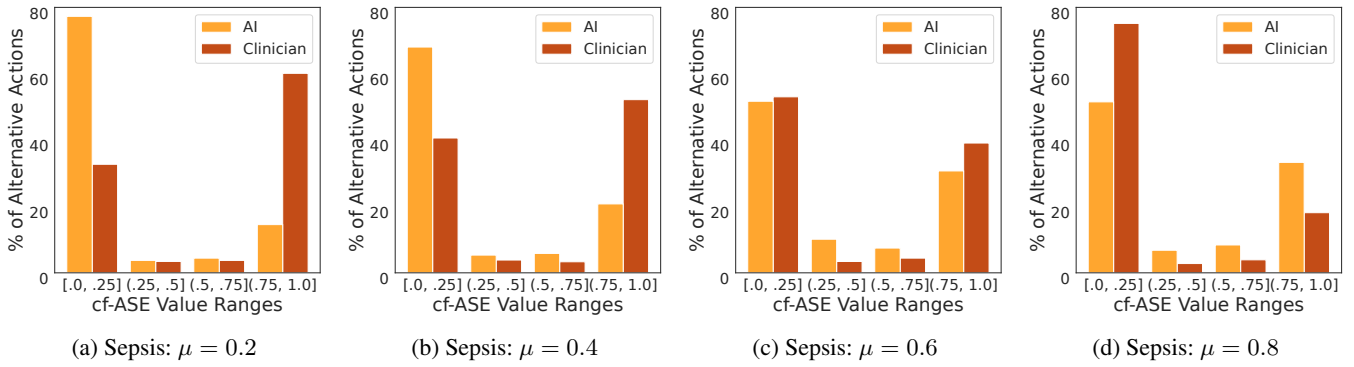


Figure 5: (Sepsis) Plots in this figure show the distribution of selected alternative actions (of both the AI and the clinician) across different ranges of cf-ASE values and for different values of trust parameter μ .

Remark 1. In the Sepsis experiments, we do not have access to the underlying causal model and instead we only have access to the simulator’s transition probabilities and the agents’ policies. Thus, for our analysis we assume that trajectories are generated by an MMDP-SCM with variables that are noise-monotonic w.r.t. a chosen total ordering. While this assumption is restrictive, we found that the results of this section are rather robust to this type of uncertainty, which is important in practice, where the assumptions we made in this paper are often violated. Plots for 5 additional randomly selected total ordering sets, which back up this claim, can be found in the Supplemental Material.

8 Discussion

To summarize, in this paper we introduce agent-specific effects, a novel causal quantity that measures the effect of an action on the outcome that propagates through other agents. To the best of our knowledge, this is the first work that looks into this problem in the context of multi-agent sequential decision making. Our theoretical contributions include results on the identifiability of this new concept, a practical algorithm to estimate it, and establishing a formal connection with prior work. In the experiments, we demonstrate the practical effectiveness of our approach in a scenario where our theoretical assumptions do not hold. Looking forward, we recognize three compelling future directions. First, we deem beneficial a characterization of the properties that agent-specific effects satisfy. Second, a natural next step of this work would be to derive an *effect attribution method* based on agent-specific effects. Finally, we plan to apply and evaluate our approach on a real-world domain.

Acknowledgements

This research was, in part, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Avin, C.; Shpitser, I.; and Pearl, J. 2005. Identifiability of path-specific effects. In *International Joint Conference on Artificial Intelligence*, 357–363.

Boutilier, C. 1996. Planning, learning and coordination in multiagent decision processes. In *Conference on Theoretical Aspects of Rationality and Knowledge*, 195–210.

Buesing, L.; Weber, T.; Zwols, Y.; Racaniere, S.; Guez, A.; Lespiau, J.-B.; and Heess, N. 2018. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*.

Carroll, M.; Chan, A.; Ashton, H.; and Krueger, D. 2023. Characterizing manipulation from AI systems. *arXiv preprint arXiv:2303.09387*.

Chiappa, S. 2019. Path-specific counterfactual fairness. In *AAAI Conference on Artificial Intelligence*, 7801–7808.

Chockler, H.; and Halpern, J. Y. 2004. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22: 93–115.

Daniel, R. M.; De Stavola, B. L.; Cousens, S. N.; and Vansteelandt, S. 2015. Causal mediation analysis with multiple mediators. *Biometrics*, 71: 1–14.

Everitt, T.; Carey, R.; Langlois, E. D.; Ortega, P. A.; and Legg, S. 2021. Agent incentives: A causal perspective. In *AAAI Conference on Artificial Intelligence*, 11487–11495.

Farquhar, S.; Carey, R.; and Everitt, T. 2022. Path-specific objectives for safer agent incentives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9529–9538.

Halpern, J. Y.; and Kleiman-Weiner, M. 2018. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *AAAI Conference on Artificial Intelligence*, 1853–1860.

Heskes, T.; Sijben, E.; Bucur, I. G.; and Claassen, T. 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in Neural Information Processing Systems*, 33: 4778–4789.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P.; Strouse, D.; Leibo, J. Z.; and De Freitas, N. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, 3040–3049.

Jung, Y.; Kasiviswanathan, S.; Tian, J.; Janzing, D.; Blöbaum, P.; and Bareinboim, E. 2022. On measuring causal

contributions via do-interventions. In *International Conference on Machine Learning*, 10476–10501.

Kämpke, T.; Radermacher, F. J.; et al. 2015. Income modeling and balancing. *Lecture Notes in Economics and Mathematical Systems*.

Komorowski, M.; Celi, L. A.; Badawi, O.; Gordon, A. C.; and Faisal, A. A. 2018. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24: 1716–1720.

Lu, C.; Huang, B.; Wang, K.; Hernández-Lobato, J. M.; Zhang, K.; and Schölkopf, B. 2020. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint arXiv:2012.09092*.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020. Explainable reinforcement learning through a causal lens. In *AAAI Conference on Artificial Intelligence*, 2493–2500.

Oberst, M.; and Sontag, D. 2019. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, 4881–4890.

Pearl, J. 1999. Probabilities Of Causation: Three Counterfactual Interpretations And Their Identification. *Synthese*, 121: 93–149.

Pearl, J. 2001. Direct and indirect effects. In *Conference on Uncertainty and Artificial Intelligence*, 411–420.

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

Pieroth, F. R.; Fitch, K.; and Belzner, L. 2022. Detecting influence structures in multi-agent reinforcement learning systems. *AAAI Workshop on Reinforcement Learning in Games*.

Richens, J.; Beard, R.; and Thompson, D. H. 2022. Counterfactual harm. *Advances in Neural Information Processing Systems*, 35: 36350–36365.

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

Singal, R.; Michailidis, G.; and Ng, H. 2021. Flow-based attribution in graphical models: A recursive shapley approach. In *International Conference on Machine Learning*, 9733–9743.

Steen, J.; Loeys, T.; Moerkerke, B.; and Vansteelandt, S. 2017. Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, 186: 184–193.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book.

Triantafyllou, S.; and Radanovic, G. 2023. Towards computationally efficient responsibility attribution in decentralized partially observable MDPs. In *International Conference on Autonomous Agents and Multiagent Systems*, 131–139.

Triantafyllou, S.; Singla, A.; and Radanovic, G. 2021. On blame attribution for accountable multi-agent sequential decision making. *Advances in Neural Information Processing Systems*, 34: 15774–15786.

Triantafyllou, S.; Singla, A.; and Radanovic, G. 2022. Actual causality and responsibility attribution in decentralized partially observable Markov decision processes. In *AAAI/ACM Conference on AI, Ethics, and Society*, 739–752.

Tsirtsis, S.; and Gomez-Rodriguez, M. 2023. Finding counterfactually optimal action sequences in continuous state spaces. *arXiv preprint arXiv:2306.03929*.

Zhang, J.; and Bareinboim, E. 2018a. Fairness in decision-making—the causal explanation formula. In *AAAI Conference on Artificial Intelligence*.

Zhang, J.; and Bareinboim, E. 2018b. Non-parametric path analysis in structural causal models. In *Conference on Uncertainty in Artificial Intelligence*.

A List of Appendices

In this section, we provide a brief description of the content provided in the appendices of the paper.

- Appendix B provides the causal graph of the MMDP-SCM from Section 2.3.
- Appendix C provides additional details on experimental setup and implementation.
- Appendix D provides additional experimental results for the Sepsis environment.
- Appendix E provides additional background needed for the proofs of our theoretical results.
- Appendix F provides additional details on fixed path-specific effects.
- Appendix G contains the proofs of Theorems 1 and 2.
- Appendix H contains the proof of Lemma 1.
- Appendix I contains the proof of Proposition 1.
- Appendix J contains the proof of Theorem 3.
- Appendix K contains the proof of Lemma 2.
- Appendix L contains the proof of Proposition 2.

B Causal Graph of MMDP-SCM

This section includes the causal graph of the MMDP-SCM described in Section 2.3. The causal graph is shown in Fig. 6.

C Additional Information on Experimental Setup and Implementation

In this section, we provide additional information on the experimental setup and implementation details.

C.1 Experimental Setup

We first provide a more detailed description of the environments and the agents’ policies used in the experiments.

Graph. We consider the graph environment depicted in Fig. 7, in which 6 agents take simultaneous actions. All agents begin at the left most node, and at each time-step move to nodes of the next column following the directed edges. Thus, each agent can select to either move *up*, *straight* or *down*, with out-of-bounds moves leading to moving straight. The agents’ goal in this environment is to be equally distributed when reaching the last column of the graph, i.e., have 2 agents per node. Agents reach the last column of the graph in 4 time-steps, when also the environment terminates.

At the initial node, all agents take a random action. At all other nodes (except the terminal ones), each agent $i \in$

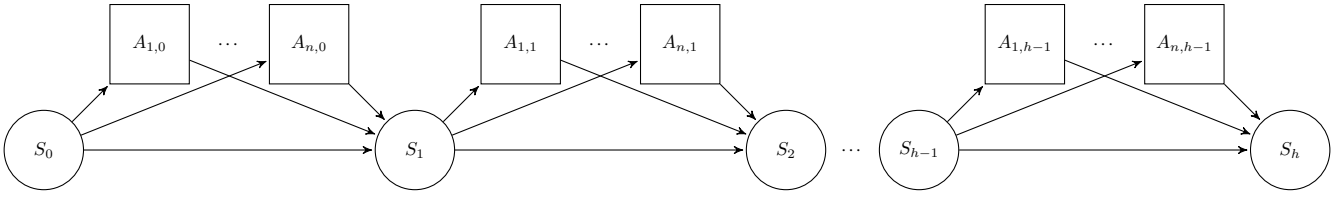


Figure 6: The causal graph of an MMDP-SCM with n agents and horizon h . Exogenous variables are omitted.

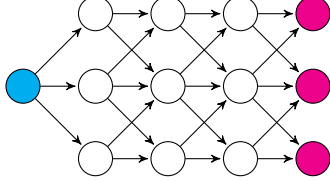


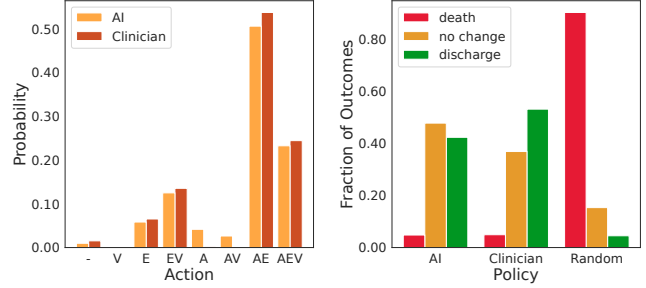
Figure 7: Graph environment. The cyan node is the initial node and magenta nodes are the terminal nodes of the graph.

$\{1, 2, 3, 4, 5, 6\}$ takes a random action with an agent-specific probability $p_i = 0.05 \cdot i$. At a node k that is occupied by a total number of agents $N_k \leq 2$, agent i moves straight with probability $1 - p_i$. Otherwise, i moves straight with probability $(1 - p_i) \cdot \frac{2}{N_k}$, and towards a row that has less than 2 agents with probability $(1 - p_i) \cdot \frac{N_k - 2}{N_k}$.¹⁰

Sepsis. The patient’s state in this environment can be described by 4 vital signs (heart rate, systolic blood pressure, percent of oxygen saturation and glucose level) as well as an additional diabetes variable present with 20% probability, influencing the fluctuations of the glucose level. Both agents can select among 8 distinct actions, each of which represents a combination of applying antibiotics (A), vasopressors (V) and mechanical ventilation (E) treatment options. As mentioned in Section 7, the clinician can additionally take a no-op action, which means that they do not override the AI’s action at that step. For more details regarding the simulator we consider in these experiments, we refer the reader to (Oberst and Sontag 2019).

The clinician’s policy is trained using the simulator’s true transition probabilities. To introduce grounds for action overriding by the clinician, the AI’s policy is trained using slightly modified transition probabilities. Namely, we increase the probability of antibiotics effectiveness and decrease the probability of vasopressors effectiveness (compared to the original simulator). This results in a policy for the AI which recommends treatments with vasopressors more scarcely and those with antibiotics more abundantly (compared to the clinician’s policy), visualized in Plot 7a. To qualitatively evaluate learned policies, we sample 1000 trajectories from the simulator and depict the observed rewards in Plot 7b. For both policies we observe similar levels of undesirable outcomes, but note that treatments prescribed by the clinician tend to lead to an earlier patient discharge, compared to those prescribed by the AI, which tend to keep

¹⁰If there are two rows with less than 2 agents where i could move to, then each of these moves has probability $(1 - p_i) \cdot \frac{N_k - 2}{2 \cdot N_k}$.



(a) Sepsis: Policy Discrepancy

(b) Sepsis: Policy Evaluation

Figure 8: (a) Probabilities that AI and clinician policies assign to different actions. (b) Observed outcomes across 1000 trajectories sampled from the simulator for AI, clinician and random policies.

the patient alive, but do not completely cure them until the end of trajectory.

We note that for a trajectory in which an undesirable outcome is realized, i.e., the patient dies at a time-step t before the completion of 20 rounds, the outcome Y is the final state of this trajectory. In other words, in our analysis we focus on the counterfactual probability that the patient would not have died at time t .

C.2 Algorithms for TCFE and cf-PSE

We now present the algorithms we use in our experiments to estimate total counterfactual effects (TCFE) and the counterfactual counterpart of path-specific effects (cf-PSE). Similar to Algorithm 1, Algorithms 2 and 3 follow the standard *abduction-action-prediction* methodology for causal inference (Pearl 2009). Furthermore, similar results to Theorem 3 can be derived for the unbiasedness of the two algorithms.

Note that for $\text{cf-PSE}_{a_i,t}^g(y|\tau)_M$ in Algorithm 3, g represents the edge-subgraph of the causal graph of M , G , with

$$\mathbf{E}(g) = \mathbf{E}(G) \setminus \bigcup_{A_{i',t'}: i' \notin \mathbf{N}, t' > t} \mathbf{E}_{-}^{A_{i',t'}}(G),$$

where $\mathbf{E}_{-}^{A_{i',t'}}(G)$ denotes the set of incoming edges of $A_{i',t'}$ in G .

C.3 Compute Architecture and Software Stack

All experiments were run on a 64bit Debian-based machine having 4x12 CPU cores clocked at 3GHz with access to 1.5TB of DDR3 1600MHz RAM. The software stack

Algorithm 2: Estimates $\text{TCFE}_{a_{i,t}}(y|\tau)_M$

Input: MMDP-SCM M , trajectory τ , action variable $A_{i,t}$, action $a_{i,t}$, outcome variable Y , outcome y , number of posterior samples H

```

1:  $h \leftarrow 0$ 
2:  $c \leftarrow 0$ 
3: while  $h < H$  do
4:    $\mathbf{u}_h \sim P(\mathbf{u}|\tau)$  {Sample noise from the posterior}
5:    $h \leftarrow h + 1$ 
6:    $y^h \sim P(Y|\mathbf{u}_h)_{M^{do(A_{i,t}:=a_{i,t})}}$  {Sample cf outcome}
7:   if  $y^h = y$  then
8:      $c \leftarrow c + 1$ 
9:   end if
10: end while
11: return  $\frac{c}{H} - \mathbb{1}(\tau(Y) = y)$ 

```

Algorithm 3: Estimates $\text{cf-PSE}_{a_{i,t}}^g(y|\tau)_M$

Input: MMDP-SCM M , trajectory τ , effect agents \mathbf{N} , action variable $A_{i,t}$, action $a_{i,t}$, outcome variable Y , outcome y , number of posterior samples H

```

1:  $h \leftarrow 0$ 
2:  $c \leftarrow 0$ 
3: while  $h < H$  do
4:    $\mathbf{u}_h \sim P(\mathbf{u}|\tau)$  {Sample noise from the posterior}
5:    $h \leftarrow h + 1$ 
6:    $I \leftarrow \{A_{i,t} := a_{i,t}\} \cup \{A_{i',t'} := \tau(A_{i',t'})\}_{i' \notin \mathbf{N}, t' > t}$ 
7:    $y^h \sim P(Y|\mathbf{u}_h)_{M^{do(I)}}$  {Sample cf outcome}
8:   if  $y^h = y$  then
9:      $c \leftarrow c + 1$ 
10:  end if
11: end while
12: return  $\frac{c}{H} - \mathbb{1}(\tau(Y) = y)$ 

```

relied on Python 3.9.13, with installed standard scientific packages for numeric calculations and visualization: NumPy (1.24.3), Pandas (2.0.3), Joblib (1.3.1), Matplotlib (3.7.1) and Seaborn (0.12.2). For learning agent’s policies in sepsis experiments, we relied on PyMDPToolbox (4.0-b3) library. The experiments are fully reproducible using the random seed 8854. The total running time of the experiments on the Graph environment is ~ 2.5 hours and of the experiments on the Sepsis environment is ~ 7.5 hours.

D Additional Experimental Results

In this section, we provide additional experimental results for the Sepsis environment. In Plots 9a-9j, we present the average *clinician-* and *AI-specific effects*, as captured by cf-ASE and the counterfactual counterpart of PSE (cf-PSE), for different levels of trust and across 5 randomly selected total ordering sets. From the plots corresponding to any of these sets of total orderings, we can draw similar conclusions to the ones we drew from Plots 4c and 4d in Section 7. Therefore, these plots indeed support our claim that the results of our experiments on Sepsis are robust to violations of our theoretical assumptions, which often happen in practice.

E Additional Theoretical Background

In this section, we provide additional background and notation needed for the proofs of our theoretical results.

Additional background on graphs. With $\mathbf{E}_-^i(G)$ or $\mathbf{E}_-^i(G)$ for short, we denote the set of incoming edges of node V^i in G . Furthermore, we use standard graph notation for relationships such as ancestors, children and descendants, e.g., $An^i(G)$ denotes the ancestor set of node V^i in G . By convention, we also assume that the ancestor and descendant sets of a node are inclusive, i.e., $V^i \in An^i(G)$ and $V^i \in De^i(G)$. We call graph g a *node-subgraph* of G if $\mathbf{V}(g) \subseteq \mathbf{V}(G)$ and $\mathbf{E}(g)$ contains all of the edges in $\mathbf{E}(G)$ connecting pairs of vertices in $\mathbf{V}(g)$.

Counterfactual formulas. We refer to realizations of a counterfactual variable, e.g., $\mathbf{Y}_w = y$ or simply y_w , as *atomic (counterfactual) formulas*. Moreover, we call *counterfactual formula* any conjunction of atomic formulas, e.g., $y_w \wedge y'_{w'}$. Let α now be a counterfactual formula defined in an SCM M , we denote with $\mathbf{V}(\alpha)$ the set of variables in \mathbf{V} that correspond to an atomic formula in α , e.g., $\mathbf{V}(y_w \wedge y'_{w'}) = \{Y\}$.

Building tools. Our proofs build on the do-calculus rules (Pearl 2009, Sec. 3.4), the exclusion and independence restrictions rules of SCMs (Pearl 2009, p. 232), and two axioms of structural counterfactuals: composition and consistency (Pearl 2009, Sec. 7.3.1).

Additional SCM notation. For some SCM M , let $\mathbf{V}' \subseteq \mathbf{V}$ and $\mathbf{u} \sim P(\mathbf{u})$. We denote with $\mathbf{V}'(\mathbf{u})$ the values of \mathbf{V}' in M given \mathbf{u} .

Monotonicity (Pearl 1999). Finally, we present here for completeness the *monotonicity* property for binary SCMs.

Definition 6 (Monotonicity). *Given a SCM M consisting of two binary variables X and Y , we say that Y is monotonic relative to X in M if for any $x_1, x_2 \in \{0, 1\}$ s.t. $\mathbb{E}[Y|x_1] \leq \mathbb{E}[Y|x_2]$ it holds that $P(Y_{x_1} = 1 \wedge Y_{x_2} = 0) = 0$.*

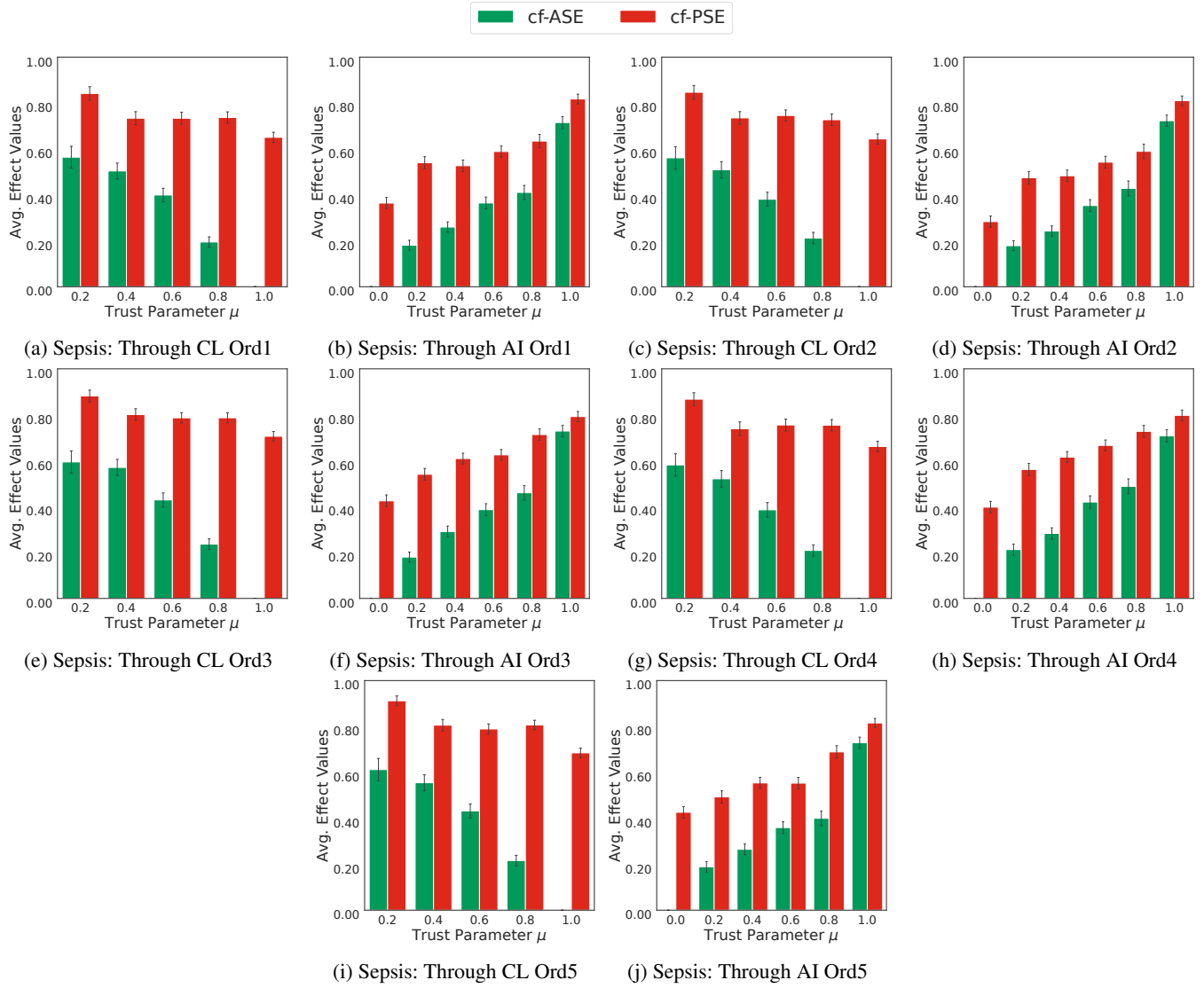


Figure 9: Plots in this figure show average values of cf-ASE and cf-PSE when computing effects that propagate through the clinician (CL) and AI, respectively, while varying trust parameter μ . Results are shown for 5 randomly selected total ordering sets (Ord1-Ord5).

F Additional Information on Fixed Path-Specific Effects

In this section, we define *counterfactual fixed path-specific effects* (cf-FPSE), i.e., the counterfactual counterpart of FPSE from Section 6. Additionally, we show that the property of noise monotonicity suffices to render cf-FPSE identifiable.

Definition 7 (cf-FPSE). *Given an SCM M with causal graph G , a realization \mathbf{v} of its observable variables \mathbf{V} , and two edge-subgraphs of G , effect subgraph g and reference subgraph g^* such that $\mathbf{E}(g) \cap \mathbf{E}(g^*) = \emptyset$, the counterfactual fixed g -specific effect of intervention $\text{do}(X := x)$ on $Y = y$, relative to reference subgraph g^* , is defined as*

$$\text{cf-FPSE}_{x,q}^{g,g^*}(y|\mathbf{v})_M = P(y_{\mathbf{v}(X)}|\mathbf{v}; M)_{M_q} - P(y|\mathbf{v})_M,$$

where q is the edge-subgraph of G with $\mathbf{E}(q) = \mathbf{E}(G) \setminus \{\mathbf{E}(g), \mathbf{E}(g^*)\}$, and $M_q = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_q, P(\mathbf{u}) \rangle$ is a modified SCM, which induces q and is formed as follows. Each function f^i in \mathcal{F} is replaced with a new function f_q^i in \mathcal{F}_q , s.t. for every $pa^i(q) \in \text{dom}\{Pa^i(q)\}$ and $\mathbf{u} \sim P(\mathbf{u})$ it holds $f_q^i(pa^i(q), u^i) = f^i(pa^i(q), pa^i(g)^e, pa^i(g^*)^*, u^i)$. Here, $pa^i(g^*)^*$ (resp. $pa^i(g)^e$) denotes the value of $Pa^i(g^*)_{\mathbf{v}(X)}$ (resp. $Pa^i(g)_x$) in M for noise \mathbf{u} .

The next lemma is similar to Lemma 2 and shows how to express cf-ASE through cf-FPSE.

Lemma 3. *Let M be an MMDP-SCM with causal graph G , \mathbf{N} be a non-empty subset of agents in M , and τ be a trajectory of M . It holds that*

$$\text{cf-ASE}_{a_i,t}^{\mathbf{N}}(y|\tau)_M = \text{cf-FPSE}_{a_i,t}^{g,g^*}(y|\tau)_M,$$

where g is the edge-subgraph of G with $\mathbf{E}(g) = \bigcup_{A_{i',t'}: i' \in \mathbf{N}, t' > t} \mathbf{E}_+^{A_{i'}, t'}(G)$ and g^* is the edge-subgraph of G with $\mathbf{E}(g^*) = \bigcup_{A_{i',t'}: i' \notin \mathbf{N}, t' > t} \mathbf{E}_+^{A_{i'}, t'}(G)$.

Proof. By Definitions 2 and 7, it follows that it suffices to show that $P(y_{\tau(A_{i,t})}|\tau; M)_{M^{do(I)}} = P(y_{\tau(A_{i,t})}|\tau; M)_{M_q}$.

Let t_Y denote the time-step of variable Y . First, we show that for any noise $\mathbf{u} \sim P(\mathbf{u}|\tau)$ the value of $S_{t_Y}(\mathbf{u})$ (or equivalently $Y(\mathbf{u})$) is the same in M_q as in $M^{do(I)}$. To do so, we use induction in the number of time-steps t' .

Base case: By construction of M_q , the value of $S_{t+1}(\mathbf{u})$ is the same in M_q as in M . Furthermore, note that the set of interventions I does not include interventions to variable S_{t+1} or to any of its ancestors. As a result, the value of $S_{t+1}(\mathbf{u})$ is the same in $M^{do(I)}$ as in M . We conclude then that the value of $S_{t+1}(\mathbf{u})$ is the same in M_q as in $M^{do(I)}$.

Induction hypothesis: We make the hypothesis that for some $t' \in [t+1, t_Y)$ the value of $S_{t'}(\mathbf{u})$ is the same in M_q as in $M^{do(I)}$.

Induction step: Under the induction hypothesis, we want to show that the value of $S_{t'+1}(\mathbf{u})$ is the same in M_q as in $M^{do(I)}$. For convenience, we slightly abuse our notation and denote as $s_{t'+1}^q$ the value of $S_{t'+1}(\mathbf{u})$ in M_q , and as $s_{t'+1}^I$ the value of $S_{t'+1}(\mathbf{u})$ in $M^{do(I)}$. From the hypothesis, it holds that the value $s_{t'}$ of $S_{t'}(\mathbf{u})$ is the same in $M^{do(I)}$ and M_q .

First, we consider the value of $S_{t'+1}(\mathbf{u})$ in $M^{do(I)}$,

$$s_{t'+1}^I = f^{S_{t'+1}}(s_{t'}, \{A_{i',t'}, a_{i,t}(\mathbf{u})\}_{i' \in \mathbf{N}}, \{\tau(A_{i',t'})\}_{i' \notin \mathbf{N}}, u^{S_{t'+1}}). \quad (2)$$

Next, we consider the value of $S_{t'+1}(\mathbf{u})$ in M_q ,

$$s_{t'+1}^q = f_q^{S_{t'+1}}(pa^{S_{t'+1}}(q), u^{S_{t'+1}}),$$

where $pa^{S_{t'+1}}(q)$ here denotes the value of $Pa^{S_{t'+1}}(q)$ in M_q given \mathbf{u} . Note that $Pa^{S_{t'+1}}(q) = S_{t'}$, and since $S_{t'}(\mathbf{u}) = s_{t'}$ in M_q , we can rewrite the equation of $s_{t'+1}^q$ as

$$s_{t'+1}^q = f_q^{S_{t'+1}}(s_{t'}, u^{S_{t'+1}}).$$

Based on the definition of function $f_q^{S_{t'+1}}$ we have that

$$s_{t'+1}^q = f^{S_{t'+1}}(s_{t'}, pa^{S_{t'+1}}(g)^e, pa^{S_{t'+1}}(g^*)^*, u^{S_{t'+1}}), \quad (3)$$

where $pa^{S_{t'+1}}(g^*)^*$ and $pa^{S_{t'+1}}(g)^e$ denote the values of $Pa^{S_{t'+1}}(g^*)_{\tau(A_{i,t})}$ and $Pa^{S_{t'+1}}(g)_{a_{i,t}}$ in M given \mathbf{u} , respectively. Since $\mathbf{u} \sim P(\mathbf{u}|\tau)$ it holds that $Z(\mathbf{u}) = \tau(Z)$ for every $Z \in \mathbf{V}$. By the consistency axiom then we have that for every $V^j \in Pa^{S_{t'+1}}(g^*)$ it holds $V_{\tau(A_{i,t})}^j(\mathbf{u}) = V^j(\mathbf{u}) = \tau(V^j)$. Moreover, from the definition of subgraph g^* it follows that $Pa^{S_{t'+1}}(g^*) = \{A_{i',t'}\}_{i' \notin \mathbf{N}}$. Eq. (3) then can be rewritten as follows

$$s_{t'+1}^q = f^{S_{t'+1}}(s_{t'}, pa^{S_{t'+1}}(g)^e, \{\tau(A_{i',t'})\}_{i' \notin \mathbf{N}}, u^{S_{t'+1}}). \quad (4)$$

Finally, from the definition of subgraph g it follows that $Pa^{S_{t'}+1}(g) = \{A_{i',t'}\}_{i' \in \mathbf{N}}$. Eq. (4) then can be rewritten as follows

$$s_{t'+1}^q = f^{S_{t'}+1}(s_{t'}, \{A_{i',t',a_{i,t}}(\mathbf{u})\}_{i' \in \mathbf{N}}, \{\tau(A_{i',t'})\}_{i' \notin \mathbf{N}}, u^{S_{t'}+1}). \quad (5)$$

From Equations (2) and (5), it follows that $s_{t'+1}^q = s_{t'+1}^I$, and hence the induction step is concluded.

Based on the induction argument above, for every $\mathbf{u} \sim P(\mathbf{u}|\tau)$ it holds that the value of $S_{t_Y}(\mathbf{u})$ or equivalently the value of $Y(\mathbf{u})$ is the same in M_q as in $M^{do(I)}$, and hence

$$P(y|\mathbf{u})_{M^{do(I)}} = P(y|\mathbf{u})_{M_q}.$$

Finally, the distribution of noise variables \mathbf{U} is the same in M , $M^{do(I)}$ and M_q , and thus we can derive the following equations

$$\begin{aligned} \int P(y|\mathbf{u})_{M^{do(I)}} \cdot P(\mathbf{u}|\tau; M)_{M^{do(I)}} d\mathbf{u} &= \int P(y|\mathbf{u})_{M_q} \cdot P(\mathbf{u}|\tau; M)_{M_q} d\mathbf{u} \Rightarrow \\ P(y|\tau; M)_{M^{do(I)}} &= P(y|\tau; M)_{M_q} \Rightarrow \\ P(y_{\tau(A_{i,t})}|\tau; M)_{M^{do(I)}} &= P(y_{\tau(A_{i,t})}|\tau; M)_{M_q}, \end{aligned}$$

where the last step follows from the consistency axiom. \square

The next result states that that noise monotonicity suffices to render cf-FPSE identifiable.

Lemma 4. *Let M be an SCM with causal graph G , g and g^* be two edge-subgraphs of G , such that $\mathbf{E}(g) \cap \mathbf{E}(g^*) = \emptyset$, and q be the edge-subgraph of G with $\mathbf{E}(q) = \mathbf{E}(G) \setminus \{\mathbf{E}(g), \mathbf{E}(g^*)\}$. Let also $\mathbf{v} \in \text{dom}\{\mathbf{V}\}$ be a realization of the observable variables in M . The counterfactual fixed g -specific effect of intervention $\text{do}(X := x)$ on $Y = y$, relative to reference subgraph g^* , is identifiable, if for every variable $Z \in \text{De}^X(G) \cap \text{An}^Y(G)$ there is a total ordering \leq_Z on $\text{dom}\{Z\}$ w.r.t. which Z is noise-monotonic in M .*

In Appendix F.1 we present three lemmas that support the proof of Lemma 4, and in Appendix F.2 we present the proof of Lemma 4.

F.1 Supporting Lemmas for the Proof of Lemma 4

Lemma 5. *Let M be an SCM with causal graph G , and \leq_X be a total ordering on $\text{dom}\{X\}$ for some $X \in \mathbf{V}$. If X is noise-monotonic in M w.r.t. \leq_X , then for any $x^1, \dots, x^k \in \text{dom}\{X\}$ and $pa^1, \dots, pa^k \in \text{dom}\{Pa^X(G)\}$ it holds that*

$$P(\wedge_{i \in \{1, \dots, k\}} x_{pa^i}^i) = \max \left(0, \min_{i \in \{1, \dots, k\}} (P(X \leq_X x^i | pa^i)) - \max_{i \in \{1, \dots, k\}} (P(X <_X x^i | pa^i)) \right).$$

Proof. Note that domain $\text{dom}\{X\}$ is finite and discrete, and hence w.l.o.g. we can assume that $\text{dom}\{X\} = \{x^1, \dots, x^m\}$, with $m \in \mathbb{N}^*$. For simplicity, we additionally assume that domain $\text{dom}\{X\}$ is sorted w.r.t. the total ordering \leq_X , i.e., $x^i <_X x^j$ for $i, j \in \{1, \dots, m\}$ such that $i < j$.

For every value pa^i of $Pa^X(G)$, we define a partition of the domain of noise variable U^X into m sets $N_{x^1}^i, \dots, N_{x^m}^i$, such that $f^X(pa^i, u^X) = x^j$ iff $u^X \in N_{x^j}^i$.

We next state a number of simple properties that are going to be useful for the rest of this proof. Let $x^j \in \text{dom}\{X\}$, $pa^i \in \text{dom}\{Pa^X(G)\}$ and $u_1^X, u_2^X \sim P(U^X)$, then the following statements hold

$$P(U^X \leq \min(u_1^X, u_2^X)) = \min(P(U^X \leq u_1^X), P(U^X \leq u_2^X)), \quad (6)$$

$$P(U^X < \max(u_1^X, u_2^X)) = \max(P(U^X < u_1^X), P(U^X < u_2^X)), \quad (7)$$

$$P(U^X \in N_{x^j}^i) = P(U^X \in [\min_u(N_{x^j}^i), \max_u(N_{x^j}^i)]), \quad (8)$$

$$\begin{aligned} P(U^X \in [u_1^X, u_2^X]) &= \begin{cases} 1 - P(U^X < u_1^X) - P(U^X > u_2^X) & \text{if } u_1^X \leq u_2^X \\ 0 & \text{otherwise} \end{cases} \\ &= \max(0, 1 - P(U^X < u_1^X) - P(U^X > u_2^X)) \\ &= \max(0, P(U^X \leq u_2^X) - P(U^X < u_1^X)). \end{aligned} \quad (9)$$

Eq. (8) follows from noise monotonicity. Moreover, by definition we have

$$P(U^X \in N_{x^j}^i) = P(X = x^j | pa^i) = P(X_{pa^i} = x^j), \quad (10)$$

$$P(U^X \leq \max_u(N_{x^j}^i)) = P(U^X \in N_{x^1}^i) + \dots + P(U^X \in N_{x^j}^i) = P(X \leq_X x^j | pa^i), \quad (11)$$

$$P(U^X < \min_u(N_{x^j}^i)) = P(U^X \in N_{x^1}^i) + \dots + P(U^X \in N_{x^{j-1}}^i) = P(X <_X x^j | pa^i). \quad (12)$$

The second step in Eq. (10) is licensed by the exogeneity assumption, while the first steps in Eq. (11) and (12) hold by the noise monotonicity assumption.

Next, we derive the equation of the lemma by using the above mentioned properties:

$$\begin{aligned}
P(\wedge_{i \in \{1, \dots, k\}} x_{pa^i}^i) &= P(U^X \in \cap_{i \in \{1, \dots, k\}} N_{x^i}^i) \\
&= P(U^X \in \cap_{i \in \{1, \dots, k\}} [\min_u(N_{x^i}^i), \max_u(N_{x^i}^i)]) \\
&= P(U^X \in [\max_{i \in \{1, \dots, k\}} (\min_u(N_{x^i}^i)), \min_{i \in \{1, \dots, k\}} (\max_u(N_{x^i}^i))]) \\
&= \max \left(0, P(U^X \leq \min_{i \in \{1, \dots, k\}} (\max_u(N_{x^i}^i))) - P(U^X < \max_{i \in \{1, \dots, k\}} (\min_u(N_{x^i}^i))) \right) \\
&= \max \left(0, \min_{i \in \{1, \dots, k\}} (P(U^X \leq \max_u(N_{x^i}^i))) - \max_{i \in \{1, \dots, k\}} (P(U^X < \min_u(N_{x^i}^i))) \right) \\
&= \max \left(0, \min_{i \in \{1, \dots, k\}} (P(X \leq_X x^i | pa^i)) - \max_{i \in \{1, \dots, k\}} (P(X <_X x^i | pa^i)) \right).
\end{aligned}$$

The first step follows from Eq. (10). The second step follows from Eq. (8). The fourth step is licensed by Eq. (9), and the fifth step by Eq. (6) and (7). Finally, the sixth step holds by Eq. (11) and (12). \square

Lemma 6. *Let M be an SCM with causal graph G , g be an edge-subgraph of G , and β be a counterfactual formula in M_g . For any $X \in \mathbf{V}$, $V^i \in De^X(g) \setminus X$ and $\mathbf{W} \subseteq \mathbf{V}$ it holds that*

$$P(\beta \wedge v_x^i | \mathbf{w}; M)_{M_g} = \sum_{pa^i(g) \in \text{dom}\{Pa^i(g)_x\}} P(\beta \wedge v_{x, pa^i(g)}^i \wedge pa^i(g)_x | \mathbf{w}; M)_{M_g}, \quad (13)$$

where $\text{dom}\{Pa^i(g)_x\}$ is equal to $\text{dom}\{Pa^i(g)\}$ if $X \notin Pa^i(g)$, and equal to the subset of $\text{dom}\{Pa^i(g)\}$ for which the value of X is always x , otherwise.

Proof.

$$\begin{aligned}
P(\beta \wedge v_x^i | \mathbf{w}; M)_{M_g} &= \sum_{pa^i(g) \in \text{dom}\{Pa^i(g)_x\}} P(\beta \wedge v_{x, pa^i(g)}^i \wedge pa^i(g)_x | \mathbf{w}; M)_{M_g} \\
&= \sum_{pa^i(g) \in \text{dom}\{Pa^i(g)_x\}} P(\beta \wedge v_{pa^i(g)}^i \wedge pa^i(g)_x | \mathbf{w}; M)_{M_g}.
\end{aligned}$$

The first step follows from the composition axiom. If $X \in Pa^i(g)$, then the second step follows from the fact that $x \wedge pa^i(g) = pa^i(g)$, which is implied by the definition of $\text{dom}\{Pa^i(g)_x\}$. Otherwise, the second step follows from the exclusion restrictions rule. \square

We next introduce Algorithm 4, which iteratively utilizes Lemma 6 to fully unroll atomic formula y_x . Lemma 7 characterizes the output of the algorithm.

Lemma 7. *Let M be an SCM with causal graph G . For two variables $X, Y \in \mathbf{V}$, let also g be an edge-subgraph of G , such that it consists only of directed paths from X to some $Z \in \mathbf{V}$ in G , and Y is part of at least one of those paths. For the input $\langle g, Y, y, X, x, \pi_g \rangle$, where π_g is a causal ordering of g , it holds that Algorithm 4 terminates after $|An^Y(g)| - 1$ iterations, and that its output α consists only of atomic formulas of the form $v_{pa^i(g)}^i$ – one for each variable $V^i \in An^Y(g) \setminus \{X\}$. Furthermore, the following equation holds*

$$P(\beta \wedge y_x | \mathbf{w}; M)_{M_g} = \sum_{\{v^i | V^i \in An^Y(g) \setminus \{X, Y\}\}} P(\beta \wedge \alpha | \mathbf{w}; M)_{M_g}, \quad (14)$$

where β is a counterfactual formula in M_g , and \mathbf{w} is a realization of $\mathbf{W} \subseteq \mathbf{V}$ in M .

Proof. We first show that at every iteration n of the algorithm, the atomic formula v_x^n is indeed part of α^n , and subsequently that the algorithm does terminate without errors after $|An^Y(g)| - 1$ iterations. To do so, we use induction in the number of iterations n .

Base case: For $n = 1$, we have that $V^1 = Y$ and $\alpha^1 = y_x$. Therefore, the statement holds for the base case $n = 1$.

Induction hypothesis: For some $k \in (1, |An^Y(g)|]$, we make the hypothesis that the statement holds for every $n < k$, i.e., we assume that the atomic formula v_x^n is part of α^n for every $n \in [1, k)$.

Algorithm 4: Iterative Process of Unrolling y_x

Input: g, Y, y, X, x, π_g
Output: counterfactual formula α

```

1:  $n \leftarrow 1$  {Number of iterations}
2:  $D^1 \leftarrow An^Y(g) \setminus \{X\}$ 
3:  $\alpha^1 \leftarrow y_x$ 
4: while  $D^n \neq \emptyset$  do
5:    $V^n \leftarrow \operatorname{argmax}_{V^i \in D^n} \{\pi_g(V^i)\}$ 
6:    $\alpha' \leftarrow \hat{\alpha}$  s.t.  $\alpha^n = \hat{\alpha} \wedge v_x^n$ 
7:    $\alpha^{n+1} \leftarrow \alpha' \wedge v_{pa^n(g)}^n \wedge pa^n(g)_x$ 
8:    $D^{n+1} \leftarrow D^n \setminus \{V^n\}$ 
9:    $n \leftarrow n + 1$ 
10: end while
11:  $\alpha \leftarrow \alpha^{|\mathcal{An}^Y(g)|}$ 
12: return  $\alpha$ 

```

Induction step: Under this hypothesis, we want to prove that v_x^k is part of α^k . Because π_g in Step 5 is a causal ordering, it holds that all variables in $Ch^k(g) \cap An^Y(g)$ must be selected by the algorithm before V^k , i.e., prior to iteration k . Let, for example, m be the first iteration at which a child of V^k is selected by the algorithm. Because of the induction hypothesis, we know that v_x^m is part of α^m . It follows then that v_x^k becomes part of α^{m+1} in Step 7 of iteration m . Since v_x^k cannot be replaced in the counterfactual formula before V^k is selected, i.e., before iteration k , we can conclude then that v_x^k is indeed part of α^k .

We next show that at every iteration n of the algorithm, the counterfactual formula α^n consists of (a) an atomic formula $v_{pa^i(g)}^i$ for each variable $V^i \in D^1 \setminus D^n$, and (b) an atomic formula v_x^i for some of the variables $V^i \in D^n$. To do so, we use induction in the number of iterations n .

(Base case) For $n = 1$, we have that $\alpha^1 = y_x$. Therefore, the statement holds for the base case $n = 1$.

(Induction hypothesis) For some $k \in [1, |\mathcal{An}^Y(g)|]$, we make the hypothesis that the statement holds for $n = k$.

(Induction step) Under this hypothesis, we want to show that the statement also holds for $n = k + 1$. From Step 7 of the algorithm, we have that $\alpha^{k+1} = \alpha' \wedge v_{pa^k(g)}^k \wedge pa^k(g)_x$, where α' is such that $\alpha^k = \alpha' \wedge v_x^k$. From the induction hypothesis and equation $\alpha^k = \alpha' \wedge v_x^k$, it follows that the counterfactual formula $\alpha' \wedge v_{pa^k(g)}^k$ consists of (a) an atomic formula $v_{pa^i(g)}^i$ for each variable $V^i \in \{V^k\} \cup D^1 \setminus D^k$, or equivalently $V^i \in D^1 \setminus D^{k+1}$, and (b) an atomic formula v_x^i for some of the variables $V^i \in D^k \setminus \{V^k\}$, or equivalently $V^i \in D^{k+1}$. Furthermore, because π_g in Step 5 is a causal ordering, it has to hold that none of the variables in $Pa^k(g)$ can be selected by the algorithm before V^k , i.e., prior to iteration k . If $X \notin Pa^k(g)$, then this implies that $Pa^k(g) \subseteq D^{k+1}$, and hence the statement holds true for $n = k + 1$. If now $X \in Pa^k(g)$, then it follows that $Pa^k(g) \setminus \{X\} \subseteq D^{k+1}$, and that x_x is part of α^{k+1} . Note, however, that x_x does not influence the truth value of α^{k+1} , since its own truth value is always 1. Hence, x_x can be removed from α^{k+1} without affecting its value. We conclude then that the statement holds true for $n = k + 1$ also when $X \in Pa^k(g)$.

Note that $\alpha = \alpha^{|\mathcal{An}^Y(g)|}$ (Step 11), and that $D^{|\mathcal{An}^Y(g)|} = \emptyset$. It follows from the last induction argument then that outcome α indeed consists of an atomic formula $v_{pa^i(g)}^i$ for each variable $V^i \in \mathcal{An}^Y(g) \setminus \{X\}$, and nothing else.

Finally, Eq. (14) simply follows from repeatedly applying Lemma 6. □

F.2 Proof of Lemma 4

Proof. Without loss of generality we assume that every edge in $\mathbf{E}(g)$ and $\mathbf{E}(g^*)$ is part of some directed path from X to Y in G . Furthermore, we will use x^* to denote the factual value $\mathbf{v}(X)$. Thus, according to Definition 7 we can write the counterfactual fixed g -specific effect as

$$\text{cf-FPSE}_{x^*, g^*}^{g, g^*}(y|\mathbf{v})_M = P(y_{x^*}|\mathbf{v}; M)_{M_q} - P(y|\mathbf{v})_M. \quad (15)$$

To identify the counterfactual fixed g -specific effect, we need to identify $P(y_{x^*}|\mathbf{v}; M)_{M_q}$ and $P(y|\mathbf{v})_M$. Note that the term $P(y|\mathbf{v})_M$ can be trivially evaluated by comparing y with $\mathbf{v}(Y)$; it is 1 if they are equal, and 0 otherwise.

We now focus on identifying $P(y_{x^*}|\mathbf{v}; M)_{M_q}$. Our first step is to express $P(y_{x^*}|\mathbf{v}; M)_{M_q}$ in terms of probabilities defined in M instead of M_q . On our second step, we will show that the expression we derived in the first step is identifiable from the observational distribution of M .

Step 1: We denote with q^d and G^d the node-subgraphs of q and G , which contain only the nodes that are parts of directed paths from X to Y in q and G , respectively. Formally, $\mathbf{V}(G^d) = De^X(G) \cap An^Y(G)$ and $\mathbf{V}(q^d) = De^X(q) \cap An^Y(q)$.

We next define the modified SCM $M_{G^d} = \langle \mathbf{U}, \mathbf{V}(G^d), \mathcal{F}_{G^d}, P(\mathbf{u}) \rangle$ which induces G^d as follows. For each variable $V^i \in \mathbf{V}(G^d)$, function f^i in \mathcal{F} is replaced with a new function $f_{G^d}^i$ in \mathcal{F}_{G^d} , such that for every $pa^i(G^d) \in \text{dom}\{Pa^i(G^d)\}$

and $u^i \sim P(u^i)$ it holds that $f_{G^d}^i(pa^i(G^d), u^i) = f^i(pa^i(G^d), \mathbf{v}(Pa^i(\overline{G^d})), u^i)$. Similarly, we define the modified SCM $M_{q^d} = \langle \mathbf{U}, \mathbf{V}(G^d), \mathcal{F}_{q^d}, P(\mathbf{u}) \rangle$ which induces q^d as follows. For each $V^i \in \mathbf{V}(q^d)$, function f_q^i in \mathcal{F}_q is replaced with a new function $f_{q^d}^i$ in \mathcal{F}_{q^d} , such that for every $pa^i(q^d) \in \text{dom}\{Pa^i(q^d)\}$ and $u^i \sim P(u^i)$ it holds that $f_{q^d}^i(pa^i(q^d), u^i) = f_q^i(pa^i(q^d), \mathbf{v}(Pa^i(\overline{G^d})), u^i)$. Note that $Pa^i(\overline{G^d}) \cap Pa^i(g) = \emptyset$ and $Pa^i(\overline{G^d}) \cap Pa^i(g^*) = \emptyset$, which follows from the fact that we assume each edge in $\mathbf{E}(g)$ and $\mathbf{E}(g^*)$ to be part of some directed path from X to Y in G . Furthermore, by the definitions of M_q and M_{q^d} it also holds that $Pa^i(q^d) \cup Pa^i(\overline{G^d}) \cup Pa^i(g) \cap Pa^i(g^*) = Pa^i(G)$.

Let set $\mathbf{S} = \cup_{V^i \in \mathbf{V}(q^d) \setminus \{X\}} Pa^i(\overline{G^d})$. We can express $P(y_{x^*} | \mathbf{v}; M)_{M_q}$ as follows

$$\begin{aligned} P(y_{x^*} | \mathbf{v}; M)_{M_q} &= P(y_{x^*}, \mathbf{v}(\mathbf{S}) | \mathbf{v}; M)_{M_q} \\ &= P(y_{x^*} | \mathbf{v}; M)_{M_{q^d}}, \end{aligned} \quad (16)$$

where the first step is licensed by exogeneity. Therefore, from Eq. (16) we have that if $P(y_{x^*} | \mathbf{v}; M)_{M_{q^d}}$ is identifiable in M then so is $P(y_{x^*} | \mathbf{v}; M)_{M_q}$.

Unroll y_{x^*} in M_{q^d} . Let α be the output of Algorithm 4 for input $\langle q^d, Y, y, X, x^*, \pi_{q^d} \rangle$, where π_{q^d} is a causal ordering of q^d . According to Lemma 7, it holds that

$$P(y_{x^*} | \mathbf{v}; M)_{M_{q^d}} = \sum_{\{v^{i,d} | V^i \in \mathbf{V}(q^d) \setminus \{X, Y\}\}} P(\alpha | \mathbf{v}; M)_{M_{q^d}}, \quad (17)$$

where counterfactual formula a consists only of atomic formulas of the form $v_{pa^i(q^d)^d}^{i,d}$ – one for each variable $V^i \in \mathbf{V}(q^d) \setminus \{X\}$. Here, we use $v^{i,d}$ to denote the value of $V_{x^*}^i$ in M_{q^d} .

Express α in M_{G^d} . Note that the counterfactual probabilities on the r.h.s. of Eq. (17) are still defined in M_{q^d} instead of M . To redefine these probabilities in M , we will need to modify α such that all implicit references to functions $f_{q^d}^i$ in M_{q^d} are removed, while its truth value is preserved. Towards that goal, we will first express α in the modified model M_{G^d} . In particular, we begin by replacing each atomic formula $v_{pa^i(q^d)^d}^{i,d}$ in α with the counterfactual formula $v_{pa^i(q^d)^d, pa^i(g)^e, pa^i(g^*)^*}^{i,d} \wedge pa^i(g)^e \wedge pa^i(g^*)^*$, where $v^{i,e}$ here denotes the value of V_x^i in M_{G^d} , and $v^{i,*}$ denotes the value of $V_{x^*}^i$ in M_{G^d} . Note also that under this new term, $v^{i,d}$ now denotes the value of $V_{x^*, pa^i(g)^e, pa^i(g^*)^*}^i$ in M_{G^d} .

Next, we group the terms of the resulting counterfactual formula as follows:

- We denote with α_1 the conjunction of all atomic formulas $v_{pa^i(q^d)^d, pa^i(g)^e, pa^i(g^*)^*}^{i,d}$. It holds that $\mathbf{V}(\alpha_1) = \mathbf{V}(q^d) \setminus \{X\}$.
- We denote with α_2 the conjunction of all atomic formulas of the form $v_x^{i,e}$. It holds that $\mathbf{V}(\alpha_2) = \cup_{V^i \in \mathbf{V}(q^d) \setminus \{X\}} Pa^i(g) \setminus \{X\}$.¹¹
- We denote with α_3 the conjunction of all atomic formulas of the form $v_{x^*}^{i,*}$. It holds that $\mathbf{V}(\alpha_3) = \cup_{V^i \in \mathbf{V}(q^d) \setminus \{X\}} Pa^i(g^*) \setminus \{X\}$.¹²

By replacing α with $\alpha_1 \wedge \alpha_2 \wedge \alpha_3$ in Eq. (17), we then have that

$$P(y_{x^*} | \mathbf{v}; M)_{M_{q^d}} = \sum_{\begin{aligned} &\{v^{i,d} | V^i \in \mathbf{V}(\alpha_1) \setminus \{Y\}\}, \\ &\{v^{i,e} | V^i \in \mathbf{V}(\alpha_2)\}, \\ &\{v^{i,*} | V^i \in \mathbf{V}(\alpha_3)\} \end{aligned}} P(\alpha_1 \wedge \alpha_2 \wedge \alpha_3 | \mathbf{v}; M)_{M_{G^d}}. \quad (18)$$

Unroll α_2 in M_{G^d} . Let α_2' be the output of Algorithm 4 for input $\langle G^d, V^i, v^{i,e}, X, x, \pi_{G^d} \rangle$, where π_{G^d} is a causal ordering of G^d and $V^i \in \mathbf{V}(\alpha_2)$. We denote with α_2' the conjunction of all such counterfactual formulas. According to Lemma 7, we can then rewrite Eq. (18) as follows

$$P(y_{x^*} | \mathbf{v}; M)_{M_{q^d}} = \sum_{\begin{aligned} &\{v^{i,d} | V^i \in \mathbf{V}(\alpha_1) \setminus \{Y\}\}, \\ &\{v^{i,e} | V^i \in \mathbf{V}(\alpha_2')\}, \\ &\{v^{i,*} | V^i \in \mathbf{V}(\alpha_3)\} \end{aligned}} P(\alpha_1 \wedge \alpha_2' \wedge \alpha_3 | \mathbf{v}; M)_{M_{G^d}}, \quad (19)$$

where counterfactual formula a_2' consists only of atomic formulas of the form $v_{pa^i(G^d)^d}^{i,d}$ – one for each variable $V^i \in \mathbf{V}(\alpha_2')$, where $\mathbf{V}(\alpha_2') = \cup_{V^i \in \mathbf{V}(\alpha_2)} An^i(G^d) \setminus \{X\}$.

Express $\alpha_1 \wedge \alpha_2' \wedge \alpha_3$ in M . To remove from α_1 , α_2' and α_3 all implicit references to functions $f_{G^d}^i$ in M_{G^d} , while preserving their truth values

¹¹We consider $X \notin \mathbf{V}(\alpha_2)$, because the atomic formula x_x has always truth value equal to 1, and hence it can be trivially removed from any counterfactual formula.

¹²We consider $X \notin \mathbf{V}(\alpha_3)$, because the atomic formula x_{x^*} has always truth value equal to 1, and hence it can be trivially removed from any counterfactual formula.

- We replace each atomic formula $v_{pa^i(q^d)^d, pa^i(g)^e, pa^i(g^*)^*}^{i,d}$ in α_1 with the new atomic formula $v_{pa^i(q^d)^d, pa^i(g)^e, pa^i(g^*)^*, \mathbf{v}(Pa^i(\overline{G^d}))}^{i,d}$. We denote with β_1 the resulting counterfactual formula. Note that it holds $\mathbf{V}(\beta_1) = \mathbf{V}(\alpha_1)$.
- We replace each atomic formula $v_{pa^i(G^d)^d}^{i,e}$ in α'_2 with the new atomic formula $v_{pa^i(G^d)^d, \mathbf{v}(Pa^i(\overline{G^d}))}^{i,e}$. We denote with β_2 the resulting counterfactual formula. Note that it holds $\mathbf{V}(\beta_2) = \mathbf{V}(\alpha'_2)$.
- We replace each atomic formula $v_{x^*}^{i,*}$ in α_3 with the new atomic formula $v_{x^*, \mathbf{v}(Pa^i(\overline{G^d}))}^{i,*}$. We denote with β_3 the resulting counterfactual formula. Note that it holds $\mathbf{V}(\beta_3) = \mathbf{V}(\alpha_3)$.

Note that counterfactual formulas β_1 , β_2 and β_3 are all defined under the original model M . Therefore, by replacing $\alpha_1 \wedge \alpha'_2 \wedge \alpha_3$ with $\beta_1 \wedge \beta_2 \wedge \beta_3$ in Eq. (19), we have that

$$P(y_{x^*} | \mathbf{v}; M)_{M_{q^d}} = \sum_{\substack{\{v^{i,d} | V^i \in \mathbf{V}(\beta_1) \setminus \{Y\}\}, \\ \{v^{i,e} | V^i \in \mathbf{V}(\beta_2)\}, \\ \{v^{i,*} | V^i \in \mathbf{V}(\beta_3)\}}} P(\beta_1 \wedge \beta_2 \wedge \beta_3 | \mathbf{v})_M. \quad (20)$$

This concludes Step 1.

Step 2: By Bayes rule, we can rewrite Eq. (20) as

$$P(y_{x^*} | \mathbf{v}; M)_{M_{q^d}} = \frac{1}{P(\mathbf{v})_M} \cdot \sum_{\substack{\{v^{i,d} | V^i \in \mathbf{V}(\beta_1) \setminus \{Y\}\}, \\ \{v^{i,e} | V^i \in \mathbf{V}(\beta_2)\}, \\ \{v^{i,*} | V^i \in \mathbf{V}(\beta_3)\}}} P(\beta_1 \wedge \beta_2 \wedge \beta_3 \wedge \mathbf{v})_M. \quad (21)$$

By consistency axiom, it holds $X = \mathbf{v}(X) \Rightarrow \mathbf{V}_{\mathbf{v}(X)} = \mathbf{V}$, and since $\mathbf{v}(X) = x^*$, we have that $\mathbf{V}_{x^*} = \mathbf{V}$. Therefore, we can rewrite Eq. (21) as follows

$$\begin{aligned} P(y_{x^*} | \mathbf{v}; M)_{M_{q^d}} &= \frac{1}{P(\mathbf{v})_M} \cdot \sum_{\substack{\{v^{i,d} | V^i \in \mathbf{V}(\beta_1) \setminus \{Y\}\}, \\ \{v^{i,e} | V^i \in \mathbf{V}(\beta_2)\}, \\ \{v^{i,*} | V^i \in \mathbf{V}(\beta_3)\}}} P(\beta_1 \wedge \beta_2 \wedge \beta_3 \wedge \mathbf{v}_{x^*})_M \\ &= \frac{1}{P(\mathbf{v})_M} \cdot \sum_{\substack{\{v^{i,d} | V^i \in \mathbf{V}(\beta_1) \setminus \{Y\}\}, \\ \{v^{i,e} | V^i \in \mathbf{V}(\beta_2)\}}} P(\beta_1 \wedge \beta_2 \wedge \mathbf{v}_{x^*})_M. \end{aligned} \quad (22)$$

The second step follows from the fact that if $v^{i,*} \neq \mathbf{v}(V^i)$, for some variable $V^i \in \mathbf{V}(\beta_3)$, then $v_{x^*}^{i,*} \wedge \mathbf{v}(V^i)_{x^*}$ always evaluates to 0.

By composition axiom, for some variable $V^i \in \mathbf{V}$ it holds that $Pa^i(G)_{x^*} = \mathbf{v}(Pa^i(G)) \Rightarrow V_{\mathbf{v}(Pa^i(G)), x^*}^i = V_{x^*}^i$. Therefore, we can rewrite Eq. (22) as follows

$$P(y_{x^*} | \mathbf{v}; M)_{M_{q^d}} = \frac{1}{P(\mathbf{v})_M} \cdot \sum_{\substack{\{v^{i,d} | V^i \in \mathbf{V}(\beta_1) \setminus \{Y\}\}, \\ \{v^{i,e} | V^i \in \mathbf{V}(\beta_2)\}}} P(\beta_1 \wedge \beta_2 \wedge \gamma)_M, \quad (23)$$

where γ is the conjunction of all atomic formulas $\mathbf{v}(V^i)_{\mathbf{v}(Pa^i(G)), x^*}$.¹³

By exclusions restrictions rule, if $X \notin Pa^i(G)$ it holds that $V_{\mathbf{v}(Pa^i(G)), x^*}^i = V_{\mathbf{v}(Pa^i(G))}^i$. Furthermore, if $X \in Pa^i(G)$ it holds $V_{\mathbf{v}(Pa^i(G)), x^*}^i = V_{\mathbf{v}(Pa^i(G))}^i$, since $\mathbf{v}(X) = x^*$. Therefore, we can rewrite Eq. (23) as follows

$$P(y_{x^*} | \mathbf{v}; M)_{M_{q^d}} = \frac{1}{P(\mathbf{v})_M} \cdot \sum_{\substack{\{v^{i,d} | V^i \in \mathbf{V}(\beta_1) \setminus \{Y\}\}, \\ \{v^{i,e} | V^i \in \mathbf{V}(\beta_2)\}}} P(\beta_1 \wedge \beta_2 \wedge \gamma')_M, \quad (24)$$

where γ' is the conjunction of all atomic formulas $\mathbf{v}(V^i)_{\mathbf{v}(Pa^i(G))}$.

¹³The order in which the composition axiom is applied here can follow any causal ordering on G .

Finally, the following equation holds by the independence restrictions rule and the exogeneity assumption.

$$\begin{aligned}
P(y_{x^*}|\mathbf{v}; M)_{M_{q^d}} &= \frac{1}{P(\mathbf{v})_M} \cdot \sum_{\substack{\{v^{i,d}|V^i \in \mathbf{V}(\beta_1) \setminus \{Y\}\}, \\ \{v^{i,e}|V^i \in \mathbf{V}(\beta_2)\}}} \\
&\quad \prod_{\{i|V^i \in \mathbf{V}(\beta_1) \cup \mathbf{V}(\beta_2)\}} P(v_{pa^i(q^d)^d, pa^i(g)^e, pa^i(g^*)^*, \mathbf{v}(Pa^i(\overline{G^d}))}^{i,d})^\wedge \\
&\quad \quad v_{pa^i(G^d)^d, \mathbf{v}(Pa^i(\overline{G^d}))}^{i,e} \wedge \mathbf{v}(V^i)_{\mathbf{v}(Pa^i(G))} M. \\
&\quad \prod_{\{i|V^i \in \mathbf{V}(\beta_1) \setminus \mathbf{V}(\beta_2)\}} P(v_{pa^i(q^d)^d, pa^i(g)^e, pa^i(g^*)^*, \mathbf{v}(Pa^i(\overline{G^d}))}^{i,d})^\wedge \\
&\quad \quad \mathbf{v}(V^i)_{\mathbf{v}(Pa^i(G))} M. \\
&\quad \prod_{\{i|V^i \in \mathbf{V}(\beta_2) \setminus \mathbf{V}(\beta_1)\}} P(v_{pa^i(G^d)^d, \mathbf{v}(Pa^i(\overline{G^d}))}^{i,e})^\wedge \mathbf{v}(V^i)_{\mathbf{v}(Pa^i(G))} M. \\
&\quad \prod_{\{i|V^i \in \mathbf{V} \setminus \{\mathbf{V}(\beta_1) \cup \mathbf{V}(\beta_2)\}\}} P(\mathbf{v}(V^i)|\mathbf{v}(Pa^i(G)))_M. \tag{25}
\end{aligned}$$

$P(\mathbf{v})_M$ can be computed from the observational distribution of M . The same also holds for the conditional probabilities $P(\mathbf{v}(V^i)|\mathbf{v}(Pa^i(G)))_M$ in the last product of the equation. Regarding the counterfactual probabilities shown on the r.h.s. of Eq. (25), according to Lemma 5, they are also identifiable under the noise monotonicity assumption. We conclude then that $P(y_{x^*}|\mathbf{v}; M)_{M_{q^d}}$ is indeed identifiable from the observational distribution of M . \square

G Proofs of Theorems 1 and 2

Theorem 1. Let M be an MMDP-SCM, and \mathbf{N} be a non-empty subset of agents in M . It holds that the \mathbf{N} -specific effect $\text{ASE}_{a_{i,t}, a_{i,t}^*}^{\mathbf{N}}(y)_M$ is non-identifiable if $t_Y > t + 1$, where t_Y denotes the time-step of variable Y .

Proof. Let G be the causal graph of M . First, we are going to prove the result for $Y = S_{t+2}$, i.e., $t_Y = t + 2$. According to Definition 1, we have that

$$\text{ASE}_{a_{i,t}, a_{i,t}^*}^{\mathbf{N}}(y)_M = P(y_{a_{i,t}^*})_{M^{do(I)}} - P(y_{a_{i,t}^*})_M,$$

where $I = \{A_{i',t+1} := A_{i',t+1, a_{i,t}^*}\}_{i' \notin \mathbf{N}} \cup \{A_{i',t+1} := A_{i',t+1, a_{i,t}}\}_{i' \in \mathbf{N}}$.

It trivially holds for $Y = S_{t+2}$ that $P(y_{a_{i,t}^*})_{M^{do(I)}} = P(y_{a_{i,t}^*})_{M^{do(I')}}$, where $I' = \{A_{i',t+1} := A_{i',t+1, a_{i,t}}\}_{i' \in \mathbf{N}}$, and hence it follows

$$\text{ASE}_{a_{i,t}, a_{i,t}^*}^{\mathbf{N}}(y)_M = P(y_{a_{i,t}^*})_{M^{do(I')}} - P(y_{a_{i,t}^*})_M. \tag{26}$$

Note that the modified SCM $M^{do(I')} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_{I'}, P(\mathbf{u}) \rangle$ is identical to M apart from the functions corresponding to variables in $\{A_{i',t+1}\}_{i' \in \mathbf{N}}$. In particular, the function $f_{I'}^{A_{j,t+1}}$ which corresponds to a variable $A_{j,t+1} \in \{A_{i',t+1}\}_{i' \in \mathbf{N}}$ in $M^{do(I')}$ can be defined as

$$\begin{aligned}
f_{I'}^{A_{j,t+1}}(Pa^{A_{j,t+1}}(G), U^{A_{j,t+1}}) &= A_{j,t+1, a_{i,t}} \\
&= f^{A_{j,t+1}}(Pa^{A_{j,t+1}}(G)_{a_{i,t}}, U^{A_{j,t+1}}) \\
&= f^{A_{j,t+1}}(S_{t+1, a_{i,t}}, U^{A_{j,t+1}}). \tag{27}
\end{aligned}$$

The second step follows from the exclusion restrictions rule. Consider now the path-specific effect

$$\begin{aligned}
\text{PSE}_{a_{i,t}, a_{i,t}^*}^g(y)_M &= P(y_{a_{i,t}^*})_{M_g} - P(y_{a_{i,t}})_{M_g} \\
&= P(y_{a_{i,t}^*})_{M_g} - P(y_{a_{i,t}})_M,
\end{aligned}$$

where g is the edge-subgraph of G that does not include the edges from S_{t+1} to nodes in $\{A_{i',t+1}\}_{i' \in \mathbf{N}}$. The second step holds because the quantity $P(y_{a_{i,t}})$ in the modified model M_g corresponds to $P(y_{a_{i,t}})$ in the original model M .

Based on Definition 4, we have that the modified SCM $M_g = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}_g, P(\mathbf{u}) \rangle$ is identical to M apart from the functions corresponding to variables in $\{A_{i',t+1}\}_{i' \in \mathbf{N}}$. In particular, the function $f_g^{A_{j,t+1}}$ which corresponds to a variable $A_{j,t+1} \in \{A_{i',t+1}\}_{i' \in \mathbf{N}}$ in M_g can be defined as

$$\begin{aligned} f_g^{A_{j,t+1}}(Pa^{A_{j,t+1}}(g), U^{A_{j,t+1}}) &= f^{A_{j,t+1}}(Pa^{A_{j,t+1}}(g), Pa^{A_{j,t+1}}(\bar{g})_{a_{i,t}}, U^{A_{j,t+1}}) \\ &= f^{A_{j,t+1}}(S_{t+1,a_{i,t}}, U^{A_{j,t+1}}). \end{aligned} \quad (28)$$

From Eq. (27) and (28), we can infer that the SCMs $M^{do(I')}$ and M_g are equivalent. Therefore, we can rewrite Eq. (26) as follows

$$\text{ASE}_{a_{i,t}, a_{i,t}^*}^{\mathbf{N}}(y)_M = P(y_{a_{i,t}^*})_{M_g} - P(y_{a_{i,t}})_M.$$

Note that quantity $P(y_{a_{i,t}^*})_M$ is trivially identifiable (the same holds for $P(y_{a_{i,t}})_M$). Thus, we can restrict our attention to the identifiability of $P(y_{a_{i,t}^*})_{M_g}$, which we can determine using the *recanting witness criterion* (Avin, Shpitser, and Pearl 2005).

The subgraph g contains the directed path $A_{i,t} \rightarrow S_{t+1} \rightarrow S_{t+2}$. As such, there is a directed path from $A_{i,t}$ to S_{t+1} in g , as well as a directed path from S_{t+1} to Y in g . Furthermore, the directed path $A_{i,t} \rightarrow S_{t+1} \rightarrow A_{j,t+1} \rightarrow S_{t+2}$, where $j \in \mathbf{N}$, does not belong to g . We can conclude then $A_{i,t}$, Y and g satisfy the recanting witness criterion, and hence $P(y_{a_{i,t}^*})_{M_g}$ is non-identifiable. Subsequently, $\text{ASE}_{a_{i,t}, a_{i,t}^*}^{\mathbf{N}}(y)_M$ is also non-identifiable.

Since we cannot identify the agent-specific effect on S_{t+2} , it follows that we cannot also identify it on any state variable $S_{t'}$ that causally depends on S_{t+2} , i.e., for any $t' > t + 2$. □

Theorem 2. *Let M be an MMDP-SCM, \mathbf{N} be a non-empty subset of agents in M , and τ be a trajectory of M . The counterfactual \mathbf{N} -specific effect $\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M$ is identifiable, if for every variable $Z \in \mathbf{V}$ there is a total ordering \leq_Z on $\text{dom}\{Z\}$ w.r.t. which Z is noise-monotonic in M .*

Proof. This result follows directly from Lemma 3 and Lemma 4. □

H Proof of Lemma 1

Lemma 1. *Let $\mathbf{X} = \{X^1, \dots, X^N\}$ represent the set of state and action variables of an MMDP, and $P(\mathbf{X})$ be the joint distribution induced by the MMDP under a joint policy π . Additionally, let \leq_1, \dots, \leq_N be total orderings on $\text{dom}\{X^1\}, \dots, \text{dom}\{X^N\}$, respectively. We can construct an MMDP-SCM $M = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$ with causal graph G , where $\mathbf{V} = \{V^1, \dots, V^N\}$, $\mathbf{U} = \{U^1, \dots, U^N\}$, $P(U^i) = \text{Uniform}[0, 1]$, and functions in \mathcal{F} are defined as*

$$f^i(Pa^i(G), U^i) = \inf_{v^i \in \text{dom}\{V^i\}} \{P(X^i \leq_i v^i | Pa^i(G))_M \geq U^i\},$$

such that $P(\mathbf{X}) = P(\mathbf{V})$ and every variable $V^i \in \mathbf{V}$ is noise-monotonic in M w.r.t. \leq_i .

Proof. We define the domain of each variable $V^i \in \mathbf{V}$ in M as $\text{dom}\{V^i\} = \text{dom}\{X^i\}$. Furthermore, we slightly abuse our notation to denote the set of variables in \mathbf{X} that correspond to the variables V^j in $Pa^i(G)$, $\mathbf{X}(Pa^i(G)) = \{X^j \in \mathbf{X} | j : V^j \in Pa^i(G)\}$.

Observe that function $f^i(Pa^i(G), U^i)$ is equivalent to the *conditional quantile function* $Q_{X^i | \mathbf{X}(Pa^i(G))=Pa^i(G)}(U^i)$, i.e., the quantile function of random variable X^i when variables in $\mathbf{X}(Pa^i(G))$ are fixed to their corresponding values in $Pa^i(G)$. We consider the following equations

$$\begin{aligned} P(V^i \leq_i v^i | Pa^i(G) = pa^i(G)) &= P(f^i(Pa^i(G), U^i) \leq_i v^i | Pa^i(G) = pa^i(G)) \\ &= P(f^i(pa^i(G), U^i) \leq_i v^i) \\ &= P(Q_{X^i | \mathbf{X}(Pa^i(G))=pa^i(G)}(U^i) \leq_i v^i) \\ &= P(U^i \leq P(X^i \leq_i v^i | \mathbf{X}(Pa^i(G)) = pa^i(G))) \\ &= P(X^i \leq_i v^i | \mathbf{X}(Pa^i(G)) = pa^i(G)). \end{aligned} \quad (29)$$

The fifth step simply follows from the fact that $U^i \sim \text{Uniform}[0, 1]$. The fourth step holds by the *reproduction* property of quantile functions (Kämpke, Radermacher et al. 2015, Theorem. 2.1), which we restate and prove for our setting next.

Lemma 8. $Q_{X^i | \mathbf{X}(Pa^i(G))=pa^i(G)}(u^i) \leq_i v^i$ iff $P(X^i \leq_i v^i | \mathbf{X}(Pa^i(G)) = pa^i(G)) \geq u^i$.

Proof.

- $P(X^i \leq_i v^i | \mathbf{X}(Pa^i(G)) = pa^i(G)) \geq u^i$ implies that

$$Q_{X^i | \mathbf{X}(Pa^i(G))=pa^i(G)}(u^i) = \inf_{x^i \in \text{dom}\{X^i\}} \{P(X^i \leq_i x^i | \mathbf{X}(Pa^i(G)) = pa^i(G)) \geq u^i\} \leq_i v^i,$$

because v^i belongs to the set over which the infimum is formed.

- Let $x_{inf}^i = \inf_{x^i \in \text{dom}\{X^i\}} \{P(X^i \leq_i x^i | \mathbf{X}(Pa^i(G)) = pa^i(G)) \geq u^i\}$. Then, $Q_{X^i | \mathbf{X}(Pa^i(G))=pa^i(G)}(u^i) \leq_i v^i$ implies that

$$v^i \geq_i x_{inf}^i.$$

Since the set $S = \{x^i \in \text{dom}\{X^i\} : P(X^i \leq_i x^i | \mathbf{X}(Pa^i(G)) = pa^i(G)) \geq u^i\}$ is finite, discrete and non-empty, the value x_{inf}^i has to belong to S . This means that $P(X^i \leq_i x_{inf}^i | \mathbf{X}(Pa^i(G)) = pa^i(G)) \geq u^i$. By monotonicity of the cumulative distribution function it has to also hold that $P(X^i \leq_i v^i | \mathbf{X}(Pa^i(G)) = pa^i(G)) \geq u^i$.

□

From Eq. (29), it follows that the MMDP-SCM M induces the transition probabilities and the initial state distribution of the MMDP, as well as the agents' joint policy π . Therefore, joint distributions $P(\mathbf{X})$ and $P(\mathbf{V})$ are indeed equal.

Next, consider two noise values $u_1^i, u_2^i \sim P(U^i)$ such that $u_1^i < u_2^i$. By monotonicity of the cumulative distribution function, for any $pa^i(G) \in \text{dom}\{Pa^i(G)\}$, it holds that

$$\begin{aligned} \inf_{v^i \in \text{dom}\{V^i\}} \{P(X^i \leq_i v^i | pa^i(G)) \geq u_1^i\} &\leq_i \inf_{v^i \in \text{dom}\{V^i\}} \{P(X^i \leq_i v^i | pa^i(G)) \geq u_2^i\} \\ \Rightarrow f^i(pa^i(G), u_1^i) &\leq_i f^i(pa^i(G), u_2^i). \end{aligned}$$

Hence, variable $V^i \in \mathbf{V}$ is indeed noise-monotonic in M w.r.t. \leq_i .

□

I Proof of Proposition 1

Proposition 1. *Let M be an SCM consisting of two binary variables X and Y with $Y = f^Y(X, U^Y)$. If Y is noise-monotonic in M w.r.t. the numerical ordering, then Y is monotonic relative to X in M . However, if Y is monotonic relative to X in M , then Y is not necessarily noise-monotonic in M w.r.t. any total ordering on $\text{dom}\{X\}$.*

Proof. We prove the first part of this lemma by contradiction. We first assume that Y is noise-monotonic in M w.r.t. the numerical total ordering \leq . Let there now be two values $x_1, x_2 \in \{0, 1\}$ such that $\mathbb{E}[Y|x_1] \leq \mathbb{E}[Y|x_2]$ and $P(Y_{x_1} = 1 \wedge Y_{x_2} = 0) > 0$ (i.e., the monotonicity assumption is violated). This means that there has to be at least one noise value u_k^Y such that $f^Y(x_1, u_k^Y) = 1$ and $f^Y(x_2, u_k^Y) = 0$. Because Y is assumed to be noise-monotonic, for any noise value $u^Y \geq u_k^Y$ it must hold that $f^Y(x_1, u^Y) = 1$, and for any noise value $u^Y < u_k^Y$ it must hold that $f^Y(x_2, u^Y) = 0$. Consequently, the following inequality holds

$$\begin{aligned} \mathbb{E}[Y|x_1] &= \int_{u^Y \sim P(U^Y)} P(u^Y) \cdot f^Y(x_1, u^Y) du^Y \\ &= \int_{u^Y \sim P(U^Y) | u^Y < u_k^Y} P(u^Y) \cdot f^Y(x_1, u^Y) du^Y + 1 + \int_{u^Y \sim P(U^Y) | u^Y > u_k^Y} P(u^Y) \cdot 1 du^Y \\ &> \int_{u^Y \sim P(U^Y) | u^Y < u_k^Y} P(u^Y) \cdot 0 du^Y + 0 + \int_{u^Y \sim P(U^Y) | u^Y > u_k^Y} P(u^Y) \cdot f^Y(x_2, u^Y) du^Y \\ &= \int_{u^Y \sim P(U^Y)} P(u^Y) \cdot f^Y(x_2, u^Y) du^Y \\ &= \mathbb{E}[Y|x_2]. \end{aligned}$$

Therefore, we have reached a contradiction.

We prove the second part of this lemma by counterexample. Assume that noise variable U^Y can only take the three values $\{1, 2, 3\}$, and that function f^Y is defined as follows

$$\begin{aligned} f^Y(X = 0, U^Y = 1) &= 1, f^Y(X = 0, U^Y = 2) = 0, f^Y(X = 0, U^Y = 3) = 0, \\ f^Y(X = 1, U^Y = 1) &= 1, f^Y(X = 1, U^Y = 2) = 0, f^Y(X = 1, U^Y = 3) = 1. \end{aligned}$$

It can be easily inferred that in this SCM, variable Y is monotonic relative to X , while it is not noise-monotonic w.r.t. the numerical total ordering or any other total ordering \leq_Y on $\{0, 1\}$.

□

J Proof of Theorem 3

Theorem 3. Let MMDP-SCMs $M = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$ and $\hat{M} = \langle \mathbf{U}, \mathbf{V}, \hat{\mathcal{F}}, \hat{P}(\mathbf{u}) \rangle$, s.t. $P(\mathbf{V})_M = P(\mathbf{V})_{\hat{M}}$ and every variable $Z \in \mathbf{V}$ is noise-monotonic in both M and \hat{M} w.r.t. the same total ordering \leq_Z on $\text{dom}\{Z\}$. It holds that the output of Algorithm 1, when it takes as input model \hat{M} , is an unbiased estimator of $\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M$.

Proof. According to Theorem 2 we have that $\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M = \text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_{\hat{M}}$. This means that the output of Algorithm 1 is the same when it takes as input either model M or model \hat{M} . Therefore, to prove Theorem 3 it would suffice to show that the output of Algorithm 1, when it takes as input MMDP-SCM M , is an unbiased estimator of $\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M$.

We begin by formally expressing the output of Algorithm 1

$$\begin{aligned} O(M, \tau, \mathbf{N}, A_{i,t}, a_{i,t}, Y, y, H) &= \frac{\sum_{\mathbf{u} \in \{\mathbf{u}_1, \dots, \mathbf{u}_H\}} P(y|\mathbf{u})_{M^{do(I(\mathbf{u}))}}}{H} - P(y|\tau)_M \\ &= \frac{\sum_{\mathbf{u} \in \{\mathbf{u}_1, \dots, \mathbf{u}_H\}} P(y|\mathbf{u})_{M^{do(I(\mathbf{u}))}} - P(y|\mathbf{u})_M}{H}, \end{aligned} \quad (30)$$

where $I(\mathbf{u})$ is defined as in line 7 of the algorithm. The second step of the equation holds because $\{\mathbf{u}_1, \dots, \mathbf{u}_H\} \sim P(\mathbf{u}|\tau)$. By Definition 2, we have that

$$\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M = P(y_{\tau(A_{i,t})}|\tau; M)_{M^{do(I)}} - P(y|\tau)_M,$$

where $I = \{A_{i',t'} := \tau(A_{i',t'})\}_{i' \notin \mathbf{N}, t' > t} \cup \{A_{i',t'} := A_{i',t',a_{i,t}}\}_{i' \in \mathbf{N}, t' > t}$.

Note that for any given noise $\mathbf{u} \sim P(\mathbf{u}|\tau)$, it holds that $A_{i',t',a_{i,t}}(\mathbf{u}) = \tau^h(A_{i',t'})$, where $\tau^h \sim P(\mathbf{V}|\mathbf{u})_{M^{do(A_{i,t}:=a_{i,t})}}$. Hence, it follows that $P(y|\mathbf{u})_{M^{do(I)}} = P(y|\mathbf{u})_{M^{do(I(\mathbf{u}))}}$. We consider next the following equations

$$\begin{aligned} \mathbb{E}_{\mathbf{u} \sim P(\mathbf{u}|\tau)}[P(y|\mathbf{u})_{M^{do(I(\mathbf{u}))}} - P(y|\mathbf{u})_M] &= \int P(y|\mathbf{u})_{M^{do(I(\mathbf{u}))}} \cdot P(\mathbf{u}|\tau) - P(y|\mathbf{u})_M \cdot P(\mathbf{u}|\tau) d\mathbf{u} \\ &= \int P(y|\mathbf{u})_{M^{do(I)}} \cdot P(\mathbf{u}|\tau) - P(y|\mathbf{u})_M \cdot P(\mathbf{u}|\tau) d\mathbf{u} \\ &= \int P(y|\mathbf{u})_{M^{do(I)}} \cdot P(\mathbf{u}|\tau; M)_{M^{do(I)}} - P(y|\mathbf{u})_M \cdot P(\mathbf{u}|\tau)_M d\mathbf{u} \\ &= P(y|\tau; M)_{M^{do(I)}} - P(y|\tau)_M \\ &= P(y_{\tau(A_{i,t})}|\tau; M)_{M^{do(I)}} - P(y|\tau)_M \\ &= \text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M, \end{aligned}$$

where the third step follows from the fact that the distribution of noise variables \mathbf{U} remains the same between M and $M^{do(I)}$. The fifth step follows from the consistency axiom. We conclude then that $O(M, \tau, \mathbf{N}, A_{i,t}, a_{i,t}, Y, y, H)$ is indeed an unbiased estimator of $\text{cf-ASE}_{a_{i,t}}^{\mathbf{N}}(y|\tau)_M$. □

K Proof of Lemma 2

Lemma 2. Let M be an MMDP-SCM with causal graph G and \mathbf{N} be a non-empty subset of agents in M . It holds that

$$\text{ASE}_{a_{i,t}, a_{i,t}^*}^{\mathbf{N}}(y)_M = \text{FPSE}_{a_{i,t}, a_{i,t}^*}^{g, g^*}(y)_M$$

where g is the edge-subgraph of G with $\mathbf{E}(g) = \bigcup_{A_{i',t'}: i' \in \mathbf{N}, t' > t} \mathbf{E}_+^{A_{i',t'}}(G)$ and g^* is the edge-subgraph of G with $\mathbf{E}(g^*) = \bigcup_{A_{i',t'}: i' \notin \mathbf{N}, t' > t} \mathbf{E}_+^{A_{i',t'}}(G)$.

Proof. By Definitions 1 and 5, it follows that it suffices to show that $P(y_{a_{i,t}^*})_{M^{do(I)}} = P(y_{a_{i,t}^*})_{M_q}$.

Let t_Y denote the time-step of variable Y . First, we show that for any noise $\mathbf{u} \sim P(\mathbf{u})$ the value of $S_{t_Y}(\mathbf{u})$ (or equivalently $Y(\mathbf{u})$) is the same in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$ as in $M^{do(I \cup A_{i,t}:=a_{i,t}^*)}$. To do so, we use induction in the number of time-steps t' .

Base case: By construction of M_q , the value of $S_{t+1}(\mathbf{u})$ is the same in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$ as in $M^{do(A_{i,t}:=a_{i,t}^*)}$. Furthermore, note that the set of interventions I does not include interventions to variable S_{t+1} or to any of its ancestors. As a result, the value of $S_{t+1}(\mathbf{u})$ is the same in $M^{do(I \cup A_{i,t}:=a_{i,t}^*)}$ as in $M^{do(A_{i,t}:=a_{i,t}^*)}$. We conclude then that the value of $S_{t+1}(\mathbf{u})$ is the same in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$ as in $M^{do(I \cup A_{i,t}:=a_{i,t}^*)}$.

Induction hypothesis: We make the hypothesis that for some $t' \in [t+1, t_Y)$ the value of $S_{t'}(\mathbf{u})$ is the same in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$ as in $M^{do(I \cup A_{i,t}:=a_{i,t}^*)}$.

Induction step: Under the induction hypothesis, we want to show that the value of $S_{t'+1}(\mathbf{u})$ is the same in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$ as in $M^{do(I \cup A_{i,t}:=a_{i,t}^*)}$. For convenience, we slightly abuse our notation and denote as $s_{t'+1}^q$ the value of $S_{t'+1}(\mathbf{u})$ in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$, and as $s_{t'+1}^I$ the value of $S_{t'+1}(\mathbf{u})$ in $M^{do(I \cup A_{i,t}:=a_{i,t}^*)}$. From the hypothesis, it holds that the value $s_{t'}$ of $S_{t'}(\mathbf{u})$ is the same in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$ as in $M^{do(I \cup A_{i,t}:=a_{i,t}^*)}$.

First, we consider the value of $S_{t'+1}(\mathbf{u})$ in $M^{do(I \cup A_{i,t}:=a_{i,t}^*)}$,

$$s_{t'+1}^I = f^{S_{t'+1}}(s_{t'}, \{A_{i',t'}, a_{i,t}(\mathbf{u})\}_{i' \in \mathbf{N}}, \{A_{i',t'}, a_{i,t}^*(\mathbf{u})\}_{i' \notin \mathbf{N}}, u^{S_{t'+1}}). \quad (31)$$

Next, we consider the value of $S_{t'+1}(\mathbf{u})$ in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$,

$$s_{t'+1}^q = f_q^{S_{t'+1}}(pa^{S_{t'+1}}(q), u^{S_{t'+1}}),$$

where $pa^{S_{t'+1}}(q)$ here denotes the value of $Pa^{S_{t'+1}}(q)$ in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$ given \mathbf{u} . Note that $Pa^{S_{t'+1}}(q) = S_{t'}$, and since $S_{t'}(\mathbf{u}) = s_{t'}$ in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$, we can rewrite the equation of $s_{t'+1}^q$ as

$$s_{t'+1}^q = f_q^{S_{t'+1}}(s_{t'}, u^{S_{t'+1}}).$$

Based on the definition of function $f_q^{S_{t'+1}}$ we have that

$$s_{t'+1}^q = f^{S_{t'+1}}(s_{t'}, pa^{S_{t'+1}}(g)^e, pa^{S_{t'+1}}(g^*)^*, u^{S_{t'+1}}), \quad (32)$$

where $pa^{S_{t'+1}}(g^*)^*$ and $pa^{S_{t'+1}}(g)^e$ denote the values of $Pa^{S_{t'+1}}(g^*)_{a_{i,t}^*}$ and $Pa^{S_{t'+1}}(g)_{a_{i,t}}$ in M given \mathbf{u} , respectively. From the definition of subgraph g^* it follows that $Pa^{S_{t'+1}}(g^*) = \{A_{i',t'}\}_{i' \notin \mathbf{N}}$. Eq. (32) then can be rewritten as follows

$$s_{t'+1}^q = f^{S_{t'+1}}(s_{t'}, pa^{S_{t'+1}}(g)^e, \{A_{i',t'}, a_{i,t}^*(\mathbf{u})\}_{i' \notin \mathbf{N}}, u^{S_{t'+1}}). \quad (33)$$

Finally, from the definition of subgraph g it follows that $Pa^{S_{t'+1}}(g) = \{A_{i',t'}\}_{i' \in \mathbf{N}}$. Eq. (33) then can be rewritten as follows

$$s_{t'+1}^q = f^{S_{t'+1}}(s_{t'}, \{A_{i',t'}, a_{i,t}(\mathbf{u})\}_{i' \in \mathbf{N}}, \{A_{i',t'}, a_{i,t}^*(\mathbf{u})\}_{i' \notin \mathbf{N}}, u^{S_{t'+1}}). \quad (34)$$

From Equations (31) and (34), it follows that $s_{t'+1}^q = s_{t'+1}^I$, and hence the induction step is concluded.

Based on the induction argument above, for every $\mathbf{u} \sim P(\mathbf{u})$ it holds that the value of $S_{t_Y}(\mathbf{u})$ or equivalently the value of $Y(\mathbf{u})$ is the same in $M_q^{do(A_{i,t}:=a_{i,t}^*)}$ as in $M^{do(I \cup A_{i,t}:=a_{i,t}^*)}$, and hence

$$P(y_{a_{i,t}^*} | \mathbf{u})_{M^{do(I)}} = P(y_{a_{i,t}^*} | \mathbf{u})_{M_q}.$$

Finally, the distribution of noise variables \mathbf{U} is the same in $M^{do(I)}$ as in M_q , and thus we can derive the following equations

$$\begin{aligned} \int P(y_{a_{i,t}^*} | \mathbf{u})_{M^{do(I)}} \cdot P(\mathbf{u})_{M^{do(I)}} d\mathbf{u} &= \int P(y_{a_{i,t}^*} | \mathbf{u})_{M_q} \cdot P(\mathbf{u})_{M_q} d\mathbf{u} \Rightarrow \\ P(y_{a_{i,t}^*})_{M^{do(I)}} &= P(y_{a_{i,t}^*})_{M_q}. \end{aligned}$$

□

L Proof of Proposition 2

Proposition 2. Given an SCM M with causal graph G , and an edge-subgraph g of G , it holds that

$$\text{PSE}_{x,x^*}^g(y)_M = \text{FPSE}_{x^*,x}^{\bar{g},g^\emptyset}(y)_M + \text{TCE}_{x,x^*}(y)_M,$$

where g^\emptyset is the edge-subgraph of G with $\mathbf{E}(g^\emptyset) = \emptyset$.

Proof. By Definition 4, we have that

$$\begin{aligned} PSE_{x,x^*}^g(y)_M &= P(y_x)_{M_g} - P(y_{x^*})_{M_g} \\ &= P(y_x)_{M_g} - P(y_{x^*})_M. \end{aligned} \quad (35)$$

The second step holds because the quantity $P(y_{a_{i,t}})$ in the modified model M_g corresponds to $P(y_{a_{i,t}})$ in the original model M . By Definition 5, we have that

$$\begin{aligned} FPSE_{x^*,x}^{\bar{g},g^\emptyset}(y)_M &= P(y_x)_{M_q} - P(y_x)_M \\ &= P(y_x)_{M_g} - P(y_x)_M. \end{aligned} \quad (36)$$

The second step follows from $\mathbf{E}(g^\emptyset) = \emptyset$. From Section 2, we have that

$$TCE_{x,x^*}(y)_M = P(y_x)_M - P(y_{x^*})_M. \quad (37)$$

By combining Equations (35), (36) and (37) we get

$$PSE_{x,x^*}^g(y)_M = FPSE_{x^*,x}^{\bar{g},g^\emptyset}(y)_M + TCE_{x,x^*}(y)_M.$$

□