Counterfactual Influence in Markov Decision Processes

Milad Kazemi* MILAD.KAZEMI@KCL.AC.UK
Jessica Lally* JESSICA.LALLY@KCL.AC.UK
Ekaterina Tishchenko EKATERINA.TISHCHENKO@KCL.AC.UK
Hana Chockler HANA.CHOCKLER@KCL.AC.UK
Nicola Paoletti NICOLA.PAOLETTI@KCL.AC.UK
Department of Informatics, King's College London, London, UK

Editors: Biwei Huang and Mathias Drton

Abstract

Our work addresses a fundamental problem in the context of counterfactual inference for Markov Decision Processes (MDPs). Given an MDP path τ, counterfactual inference allows us to derive counterfactual paths τ' describing *what-if* versions of τ obtained under different action sequences than those observed in τ. However, as the counterfactual states and actions deviate from the observed ones over time, *the observation τ may no longer influence the counterfactual world*, meaning that the analysis is no longer tailored to the individual observation, resulting in interventional outcomes rather than counterfactual ones. This issue specifically affects the popular Gumbel-max structural causal model used for MDP counterfactuals, and yet, it has remained overlooked until now. In this work, we introduce a formal characterisation of influence based on comparing counterfactual and interventional distributions. We devise an algorithm to construct counterfactual models that automatically satisfy influence constraints. Leveraging such models, we derive counterfactual policies that are not just optimal for a given reward structure but also remain tailored to the observed path. Even though there is an unavoidable trade-off between policy optimality and strength of influence constraints, our experiments demonstrate that it is possible to derive (near-)optimal policies while remaining under the influence of the observation.

Keywords: Counterfactual Inference, Markov Decision Processes

1. Introduction

Counterfactual inference allows us to reason hypothetically about what effect changing an action or condition in the past would have on a given observation, e.g., *"What would the patient's condition be, had we used treatment Y instead of treatment X?"*. This allows us to evaluate and optimise sequences of actions, by identifying how the outcome could have been improved by changing some action(s) along the observed path. This optimised counterfactual path can be seen as a *counterfactual explanation* for how the decision-making process could be improved. Markov Decision Processes (MDPs) are particularly useful for modelling real-world decision-making processes under uncertainty, so counterfactual inference can be used to generate counterfactual explanations for improving a given policy. This research area aligns with the increasing use of machine learning in healthcare, specifically in supporting clinical decision-making with the use of AI models (Verma et al., 2020; Guidotti, 2022; Tsirtsis et al., 2021). In particular, we focus on discrete-state MDPs, a fundamental computation model used in e.g., clinical decision-making (Bennett and Hauser, 2013; Shifrin and Siegelmann,

* Equal contribution.

2020; Patrick and Begen, 2011), games (Hafner et al., 2023), and planning tasks (Natarajan and Kolobov, 2022).

However, one previously neglected issue with performing counterfactual inference in sequential decision processes (which is not an issue in single-step scenarios) is that, after some number of steps, the observation may no longer inform the counterfactual transition probabilities of the counterfactual path. Consider an observed path $\tau = (s_1, a_1, s_2, a_2, s_3)$. Formally, each s_{t+1} arises from a structural causal model (SCM), $s_{t+1} \sim f(s_t, a_t, \mathbf{U})$, where $\mathbf{U}$ are the (prior) random exogenous factors and f is a deterministic function. We can apply counterfactual inference to determine, for example, which states we would have reached if we had instead performed action $a'_1 \neq a_1$ at time $t = 1$, and $a'_2 \neq a_2$ at $t = 2$. This involves deriving the posterior $\mathbf{U}' = \mathbf{U} \mid \tau$ and predicting $s'_2 \sim f(s_1, a'_1, \mathbf{U}')$ and $s'_3 \sim f(s'_2, a'_2, \mathbf{U}')$. At $t = 1$, the observed transition informs the counterfactual transition probabilities (e.g., as we observed s_2, this state would be more likely in the counterfactual world). However, if the counterfactual state s'_2 diverges from the observed s_2 so that the distribution of the counterfactual outcome s'_3 is not at all influenced by the observation τ, the counterfactual probability for the transition $s'_2, a'_2 \to s'_3$ will be equal to its prior (interventional) probability, as if no observation occurred. This way, the counterfactual analysis is no longer tailored to the individual observation.

In this paper, we formalise this notion of *counterfactual influence* in MDPs. In particular, we say that an observed path influences the counterfactual world if the interventional probabilities (i.e., the transition probabilities of the nominal MDP under the counterfactual policy) and the corresponding counterfactual probabilities are not identical. Using this concept, we can ensure our policy analysis is based on counterfactual paths that are sufficiently informed by the observation.

Motivating Example To illustrate why influence is important for counterfactual analysis, consider Figure 1. This depicts a subset of the state space for the Sepsis MDP (Oberst and Sontag, 2019), which simulates trajectories of sepsis patients. Each dot represents a unique state (which are unordered), and each row corresponds to a time step in an MDP trajectory. When treating sepsis, it is important to consider whether the patient has diabetes, as diabetic patients may respond differently to treatment (e.g., with more varying blood glucose levels) than non-diabetic patients. Figure 1 illustrates this difference in terms of how often both groups visit each MDP state: the spectrum from red to blue represents how frequently states are reached by diabetic patients (in red), compared with the whole population (in blue), and the intensity of the colour represents how frequently states are visited by both groups.

Given an observed trajectory of a diabetic patient (the black line in Figure 1), we can use counterfactual inference to find the optimal counterfactual path (with the highest cumulative reward), to determine how their treatment could have been improved. The optimal counterfactual path (blue line) loses the information from the observation that the patient is diabetic, as it visits states (in blue) that are typical for the population on average but unlikely for diabetic patients. However, by constraining the optimal counterfactual path to retain influence from the observation, we obtain a counterfactual path (red line) more likely for a diabetic patient. Importantly, both counterfactual paths significantly improve on the observed path, as the patient has fewer vitals out of range at most time steps. Still, only the influenced path is tailored to the observation, albeit slightly sub-optimal compared to the unconstrained path. Also, if we derive counterfactual policies based on areas of the counterfactual MDP that are not informed/influenced by the observation, we risk identifying policies that would be optimal for the general population and not optimal for the observed (diabetic) patient. In Appendix A, we provide simulation evidence of this effect in the Sepsis MDP.

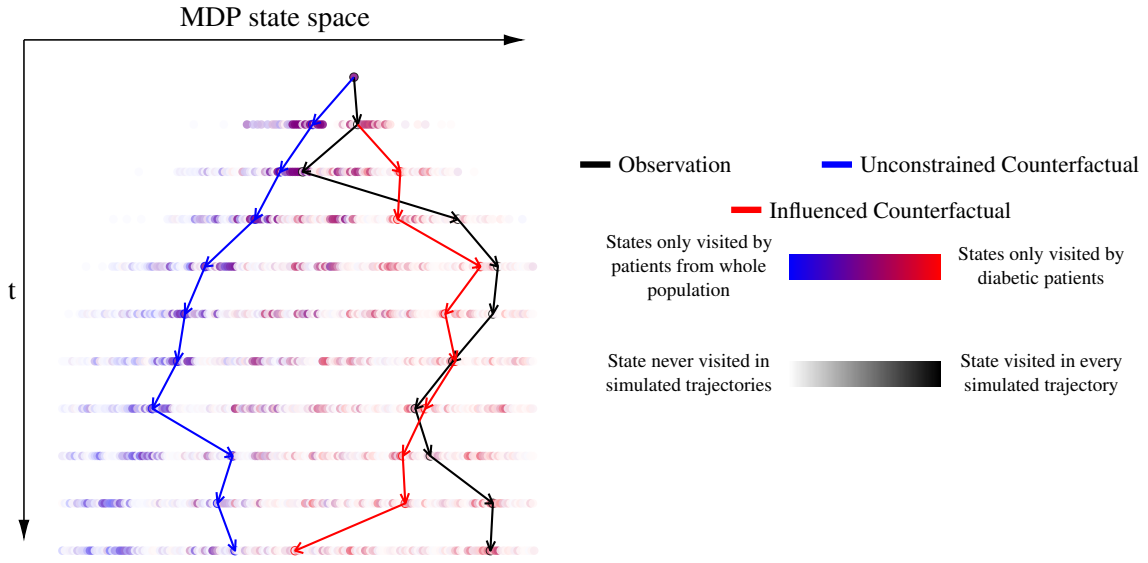


Figure 1: Subset of the state space of the Sepsis MDP. The spectrum from blue to red represents how frequently the state appears in simulated paths of diabetic patients (in red) vs. the whole population (in blue). The intensity of the colour represents how frequently the states are visited in the simulated paths. The black line is an observed trajectory for a diabetic patient; the blue line is the unconstrained counterfactual generated for that path, and, in red, the influence-constrained counterfactual path. The unconstrained counterfactual path diverges further from the the observation than the influenced counterfactual, as the observation and influenced counterfactual reach the same state at multiple timesteps ($t = 1$ and $t = 6$), and the states along both paths are shaded with a similar hue and intensity of red, indicating these paths have comparable (high) likelihoods of occurring in diabetic patients (vs. the general population), unlike the unconstrained counterfactual which visits a completely disjoint set of states.

Although this is a simplified example (as the diabetic status of a patient would be known, and their treatment can be adjusted accordingly), there may be situations where there is an underlying difference between sub-populations (e.g., patients with different genotypes) which is unknown. In these cases, it is beneficial to restrict counterfactual paths to ensure they remain tailored to the observations.

Contributions. Leveraging our formal definition of influence, we propose an approach to derive counterfactual policies that are not only optimal for a given reward structure, but also maintain a degree of influence from the observed trajectory. This ensures that the counterfactuals remain tailored to the given observation. We build on recent work by Oberst and Sontag (2019) which introduces Gumbel-max SCMs, a class of causal models for counterfactual inference of discrete-state MDPs, and Tsirtsis et al. (2021), which leverages Gumbel-max SCMs to derive an alternative sequence of actions that optimises the counterfactual outcome with a limit on the number of actions that can be changed. However, these and other existing papers on counterfactual analysis of MDPs, e.g. (Buesing et al., 2018; Lu et al., 2020; Kazemi et al., 2025), ignore the issue of influence, resulting in counterfactuals that may be (erroneously) disconnected from the observation. To the

best of our knowledge, we are the first to identify this issue of counterfactual influence in sequential decision-making processes, and our paper is intended to support the expanding field of work applying counterfactual inference to MDPs by addressing this problem. A thorough discussion of related work can be found in Appendix B.

Procedurally, we build a counterfactual MDP that inherently satisfies influence constraints, through a polynomial-time algorithm that prunes the non-influenced transitions from the counterfactual MDP. We note that imposing this constraint may be too stringent, and results in counterfactual MDPs that are close or equal to the observed path. Therefore, we further extend the notion of influence to encompass multiple steps, so that influence constraints must hold at least once (as opposed to always) over paths of a given length. In this way, we can arbitrarily relax influence constraints, allowing for larger deviations from the observation, to favour optimality of the counterfactual policy.

We validate our approach on a Grid World model and two health-related case studies: an epidemic model and a sepsis model. We evaluate how relaxing the influence constraint impacts the derived optimal policy, and find that our approach results in counterfactual paths that are optimal or near-optimal while remaining influenced by the observed path. This results in more informative counterfactual explanations for improving a given policy, as these explanations are guaranteed to be tailored toward the given observation.

2. Preliminaries

MDPs are a class of stochastic models to describe sequential decision-making processes. In an MDP $\mathcal{M}$, at each step t, an agent in state s_t performs some action a_t determined by a policy π, ending up in state $s_{t+1} \sim P_{\mathcal{M}}(s \mid s_t, a_t)$. The agent receives some reward $R(s_t, a_t)$ for performing a_t at s_t. Formally, an MDP is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P_{\mathcal{M}}, P_I, R)$ where $\mathcal{S}$ is the discrete *state space*, $\mathcal{A}$ is the set of *actions*, $P_{\mathcal{M}} : (\mathcal{S} \times \mathcal{A} \times \mathcal{S}) \to [0, 1]$ is the *transition probability* function, $P_I : \mathcal{S} \to [0, 1]$ is the *initial state distribution*, and $R : (\mathcal{S} \times \mathcal{A}) \to \mathbb{R}$ is the *reward function*. A (deterministic) *policy* π for $\mathcal{M}$ is a function $\pi : \mathcal{S} \to \mathcal{A}$. A path τ of $\mathcal{M}$ under policy π, denoted $\tau \sim \mathcal{M}(\pi)$, is a sequence $\tau = (s_0, a_0), (s_1, a_1), \ldots, (s_{T-1}, a_{T-1})$ where $T = |\tau|$ is the path length, $P_I(s_0) > 0$, $a_t = \pi(s_t)$ for all $t = 0, \ldots, T-1$, and $P_{\mathcal{M}}(s_{t+1} \mid s_t, a_t) > 0$ for all $t = 0, \ldots, T-2$.

While the MDP characterisation helps make predictions about future states and design action policies, it is not sufficient to make counterfactual predictions. For this, we require *structural causal models (SCMs)* (Pearl, 2009). Formally, an SCM is a tuple $\mathcal{C} = (\mathbf{U}, \mathbf{V}, \mathcal{F} = \{f_V\}_{V \in \mathbf{V}}, P(\mathbf{U}))$, where $\mathbf{U}$ is a set of mutually independent *exogenous variables* with $P(\mathbf{U})$ being its distribution, and $\mathbf{V}$ is a set of *endogenous variables*. The value of each $V \in \mathbf{V}$ is determined by a function $V = f_V(\mathbf{PA}_V, U_V)$ where $\mathbf{PA}_V \subseteq \mathbf{V}$ is the set of direct causes of V and $U_V \in \mathbf{U}$.

Exogenous variables are unobserved and act as the source of randomness in the model. For a fixed realisation $\mathbf{u} \sim P(\mathbf{U})$, i.e., a concrete unfolding of the system's randomness, the values of $\mathbf{V}$ become deterministic, as they are uniquely determined by $\mathbf{u}$ and $\mathcal{F}$. The value $\mathbf{u}$ is also called *context*. With SCMs, one can establish the causal effect of some input variable X on some outcome Y by evaluating Y after applying an *intervention* $X \leftarrow x$, i.e., after replacing the RHS of $X = f_X(\mathbf{PA}_X, U_X)$ with x. We denote the resulting *interventional distribution* by $P_{\mathcal{C}[x]}(Y)$ (also written as $P(Y \mid do(x))$ in Pearl's *do* notation).

Upon observing a realisation $\mathbf{v}$ of the SCM variables $\mathbf{V}$, counterfactuals answer the following question: *what would have been the value of variable Y under intervention $X \leftarrow x$, given that we observed $\mathbf{v}$?* This corresponds to evaluating $\mathbf{V}$ in a hypothetical world characterised by the same

context that generated the observation $\mathbf{v}$ but under a different causal process (i.e., $\mathcal{C}[x]$ instead of $\mathcal{C}$). Computing counterfactuals first requires deriving the context that led to the observation, i.e., $\mathbf{U} \mid \mathbf{v}$, then evaluating the outcome under the counterfactual model obtained from $\mathcal{C}[x]$ by replacing $P(\mathbf{U})$ with the inferred $P(\mathbf{U} \mid \mathbf{v})$. Since each observation $\mathbf{v}$ can be seen as a deterministic function of a particular context $\mathbf{u}$, then in theory, the counterfactual outcome should be deterministic too. However, as we will see, $\mathbf{u}$ cannot be always uniquely identified from $\mathbf{v}$, resulting in a (non-Dirac) posterior distribution $P(\mathbf{U} \mid \mathbf{v})$ and a stochastic counterfactual outcome.

2.1. SCM-based Encoding of MDPs

To enable counterfactual reasoning in MDPs, Oberst and Sontag (2019) proposed the following SCM-based encoding. For an MDP $\mathcal{M}$, policy π, and horizon T, we define an SCM over variables $\{S_t, A_t\}_{t=0}^{T-1}$ (which describes paths of length T induced by $\mathcal{M}$ and π) with the following structural equations:

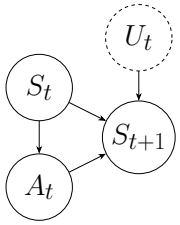


Figure 2: MDP causal graph

$$S_{t+1} = f(S_t, A_t, U_t); \;\; A_t = \pi(S_t); \;\; S_0 = f_0(U_0) \tag{1}$$

This encoding allows us to compute counterfactuals, but it is not obvious how to define f (and f_0) when dealing with categorical variables (arising from the MDP's discrete state space). To this purpose, Oberst and Sontag (2019) introduced *Gumbel-Max SCMs*, expressed as

$$S_{t+1} = f(S_t, A_t, U_t = (G_{s,t})_{s\in\mathcal{S}}) = \arg\max_{s\in\mathcal{S}} \{\log\left(P_{\mathcal{M}}(s \mid S_t, A_t)\right) + G_{s,t}\} \tag{2}$$

where for each $s \in \mathcal{S}$ and $t = 0, \ldots, T-1$, $G_{s,t}$ follows a Gumbel distribution. This approach is grounded in the Gumbel-Max trick, which shows that sampling from a categorical distribution with k categories is equivalent to sampling k copies $g_0, \ldots, g_k$ of a standard Gumbel and then determining the outcome as $\arg\max_j \{\log\left(P(Y=j)\right) + g_j\}$. Notably, the Gumbel-Max SCM encoding possesses a desirable feature called *counterfactual stability*, which, informally, states that a counterfactual outcome remains equal to the observed outcome unless the intervention increases the relative probability of an alternative outcome[1]. See (Oberst and Sontag, 2019) for more details.

We note that, although the Gumbel-Max SCM is not the only SCM capable of expressing categorical distributions (Zhang et al., 2022), such as those in MDPs, it remains the most popular causal model for MDPs (Lorberbom et al., 2021; Benz and Rodriguez, 2022; Noorbakhsh and Rodriguez, 2022; Killian et al., 2022; Zhu et al., 2020; Tsirtsis and Rodriguez, 2023).

Counterfactual inference. Given an MDP path $\tau = (s_0, a_0), \ldots, (s_{T-1}, a_{T-1})$, counterfactual inference requires identifying the values of the Gumbel exogenous variables that align with τ, i.e., calculate $\mathbf{G}' = (G_{s,t})_{s\in\mathcal{S}}^{t=0,\ldots,T-1} \mid \tau$. Because MDPs are Markovian, we can infer the Gumbel values for each observed transition in τ independently. However, because the mechanism of (2) is non-invertible (i.e., for given s_t and a_t, multiple sets of $(G_{s,t})_{s\in\mathcal{S}}$ can result in the same s_{t+1}), we cannot uniquely identify the Gumbel values. Instead, as proposed by Oberst and Sontag (2019), we can achieve (approximate) posterior inference of $P((G_{s,t})_{s\in\mathcal{S}} \mid s_t, a_t, s_{t+1})$ through *rejection sampling*: we draw samples from the prior $(g_{s,t})_{s\in\mathcal{S}} \sim P((G_{s,t})_{s\in\mathcal{S}})$ and discard all instances where $f(s_t, a_t, (g_{s,t})_{s\in\mathcal{S}}) \neq s_{t+1}$. This can also be implemented more efficiently using the top-down Gumbel sampling approach described in (Maddison et al., 2014).

1. Counterfactual stability doesn't hold for instance if we encode the MDP using the inverse CDF trick.

2.2. Counterfactual MDP and Optimal Policies

Given any MDP $\mathcal{M}$ (with known transition probabilities $P_{\mathcal{M}}$) and observed path τ, we can define a corresponding *counterfactual MDP* $\mathcal{M}^{\tau}$ which captures the counterfactual probabilities at any choice of state s and action a. $\mathcal{M}^{\tau}$ will be a non-stationary MDP with the same state space $\mathcal{S}$, action space $\mathcal{A}$, and reward structure R as $\mathcal{M}$. Its initial state distribution P_I^{τ} assigns probability 1 to s_0, the first state of τ; and its transition probabilities directly follow from the SCM (2) and are defined, for $t = 0, \ldots, T-1$ and $\forall s' \in \mathcal{S}$, as

$$P_{\mathcal{M},t,\tau}(s' \mid s, a) = P\left(s' = \arg\max_{q \in \mathcal{S}} \left\{\log\left(P_{\mathcal{M}}(q \mid s, a)\right) + G'_{q,t}\right\}\right)$$
$$\approx \frac{1}{N}\sum_{j=0}^{N} \mathbb{1}\left(s' = \arg\max_{q \in \mathcal{S}} \left\{\log\left(P_{\mathcal{M}}(q \mid s, a)\right) + G'^{(j)}_{q,t}\right\}\right) \quad (3)$$

where we sample N values $G'^{(j)}_{q,t}$ from the true posterior distribution $G'_{q,t}$ using either the rejection sampling or top-down sampling approach. The indicator function $\mathbb{1}(\mathbb{X})$ takes the value 1 if the condition $\mathbb{X}$ is satisfied and 0 otherwise. $\mathcal{M}^{\tau}$ can be directly solved to derive optimal counterfactual policies, as done in (Tsirtsis et al., 2021). In that paper, the authors are concerned with finding optimal action sequences that deviate from the observed path by at most m actions. We call this an *m-CF policy*:

Definition 1 (m-CF policy (Tsirtsis et al., 2021)) *Let τ be a path of an MDP $\mathcal{M}$ of length T, and let $\mathcal{M}^{\tau}$ be the corresponding counterfactual MDP. For a given $m \leq T$, an m-*CF policy *π^* is one that maximises the value $V_{\tau}(\pi) = \mathbb{E}_{\tau' \sim \mathcal{M}^{\tau}(\pi)}\left[\sum_{t=0}^{T-1} R(s'_t, a'_t)\right]$ under the condition that any counterfactual path τ' induced by $\mathcal{M}^{\tau}$ and π^* satisfies $\sum_{t=0}^{T-1} \mathbb{1}(a_t \neq a'_t) \leq m$.*

To derive m-CF policies, Tsirtsis et al. (2021) employ a polynomial-time dynamic programming algorithm that keeps track of the number of action changes between the observed path τ and the counterfactual one τ'. However, their approach has two main shortcomings. First, τ' may diverge from τ, visiting different state sequences where it may not be sensible to impose the same observed action. Second, and most important, because of this divergence, it is possible to reach a state in τ' where the observed path τ bears no longer influence, meaning that we are no longer computing counterfactuals but interventional outcomes. In the next sections, we characterise and solve the latter issue.

3. Theoretical Framework

This section introduces a formal notion of *influence*, to describe how an observed path τ affects the counterfactual world. For counterfactual MDPs derived using Gumbel-max SCMs (see (3)), the counterfactual distribution $P_{\mathcal{M},t,\tau}$ will be identical to the nominal/interventional distribution $P_{\mathcal{M}}$ when there is no influence. Before formally defining influence, we first derive a sufficient condition for these two distributions to be equal when using Gumbel-max SCMs:

Proposition 2 *Let τ be a path of an MDP $\mathcal{M}$ of length T, and let $\mathcal{M}^\tau$ be the corresponding counterfactual MDP. Given a time $t < T$ and counterfactual state s'_t and action a'_t in $\mathcal{M}^\tau$, then if $P_\mathcal{M}(\cdot \mid s_t, a_t)$ and $P_\mathcal{M}(\cdot \mid s'_t, a'_t)$ have disjoint support, then the counterfactual distribution $P_{\mathcal{M},t,\tau}(\cdot \mid s'_t, a'_t)$ is identical to the interventional distribution $P_\mathcal{M}(\cdot \mid s'_t, a'_t)$.*

Proof At time t, if the distributions $P_\mathcal{M}(\cdot \mid s_t, a_t)$ and $P_\mathcal{M}(\cdot \mid s'_t, a'_t)$ have disjoint supports then the posterior Gumbel distribution relative to the possible next states from (s'_t, a'_t) remains the same as the prior Gumbel distribution for those states. This is because the states in the support of $P(\cdot \mid s'_t, a'_t)$ cannot be reached from (s_t, a_t) in the factual world (i.e., in the original MDP), hence, no observation can change their prior. As a consequence, by (2) and (3), the counterfactual probabilities for those states remain equal to the probabilities in the original MDP. ■

With this proposition, we can define influence by checking a simple condition on the transition probabilities of the original MDP. Importantly, this condition is precise because the probabilities $P_\mathcal{M}$ are known. On the other hand, the equality $P_{\mathcal{M},t,\tau}(\cdot \mid s'_t, a'_t) = P_\mathcal{M}(\cdot \mid s'_t, a'_t)$ cannot be established precisely because of the sampling error introduced in $P_{\mathcal{M},t,\tau}$ by the Gumbel posterior inference[2].

Definition 3 (1-step influence) *Let τ be a path of an MDP $\mathcal{M}$ of length T, and let $\mathcal{M}^\tau$ be the corresponding counterfactual MDP. Given a time $t < T$ and counterfactual state s'_t and action a'_t in $\mathcal{M}^\tau$, we say that τ exerts an immediate (1- step) influence on s'_t and a'_t at time t if and only if the supports of the distributions $P_\mathcal{M}(\cdot \mid s'_t, a'_t)$ and $P_\mathcal{M}(\cdot \mid s_t, a_t)$ are not disjoint.*

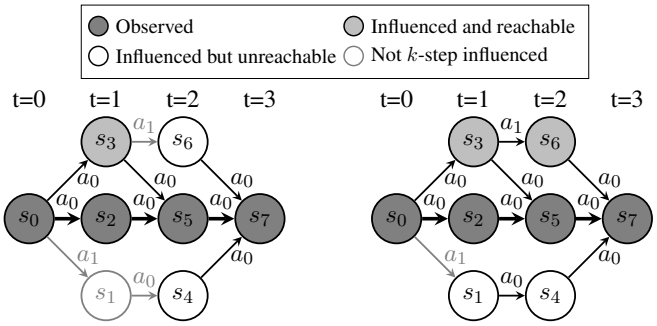
While imposing influence constraints is clearly desirable, the above notion of 1-step influence may be too strict and we may not be able to deviate much, or at all, from the observed path. To overcome this potential limitation, we relax and generalise the notion of 1-step influence to encompass multiple steps, such that influence constraints must hold at least once in a counterfactual path, and not at every step[3].

Definition 4 (k-step influence) *Let $\mathcal{M}$, T, τ, and $\mathcal{M}^\tau$ be as in Definition 3. Given a time $t < T$, horizon k, and counterfactual state s'_t and action a'_t, if $t + k \leq T$ we say that τ exerts a k-step influence on s'_t and a'_t at time t if there exists a path τ' of $\mathcal{M}^\tau$ of length k starting in (s'_t, a'_t) and such that τ exerts a 1-step influence on at least one state of τ'. If $t + k > T$, τ always exerts a k-step influence on s'_t and a'_t at time t.*

Remark 5 *In the above definition, when $t + k > T$, the length T of the observed path is not sufficient to determine k-step influence, and so we make the conservative assumption that influence holds in such cases. In particular, for $k \geq T + 1$, we have that the observed path τ exerts a k-step influence on any counterfactual state at any time t (because when $k \geq T + 1$, then $t + k > T$ for any t).*

2. It is possible to define other notions of influence that are more quantitative, e.g., based on the statistical distance between the transition probabilities. This is discussed in more detail in Appendix C.
3. The issue of (lack of) influence is especially pronounced in sequential processes (that over time, may drift away from the observed path). However, this problem can also in a static (i.e., 1-step) setting. Consider an action a' that, when applied to the initial observed state s_0 (rather than the observed action a_0) leads to a next state distribution $P(\cdot|s_0, a')$ that has disjoint support w.r.t. $P(\cdot \mid s_0, a_0)$. By Proposition 2, this implies that the counterfactual distribution will be equal to the interventional one, i.e., the observation (s_0, a_0, s_1) doesn't inform the counterfactual outcome for (s_0, a').

Figures 3(*a*) and 3(*b*) depict a rollout of a toy counterfactual MDP with two possible actions, a_0 and a_1, at states s_0 and s_3, and only action a_0 at the rest of the states. Given the observed path $\tau = [(s_0, a_0), (s_2, a_0), (s_5, a_0), (s_7, a_0)]$, we want to find the influence-constrained counterfactual MDP given $k = 1$ and $k = 2$. When $k = 1$ (Figure 3(*a*)), (s_1, a_0) and (s_3, a_1) are not influenced at $t = 1$, because they have disjoint support with the observed state-action pair (s_2, a_0). For the opposite reason, we have that (s_2, a_0) and (s_3, a_0) are influenced at $t = 1$. (s_4, a_0), (s_5, a_0) and (s_6, a_0) are all influenced at $t = 2$, as these have overlapping support with the observed pair (s_5, a_0). However, even though (s_4, a_0) and (s_6, a_0) are influenced, they cannot be reached from any influenced state-action pairs, so are also removed from the influence-constrained counterfactual MDP: we say they are "influenced but unreachable". Figure 3(*b*) depicts the case of 2-step influence. We note that (s_6, a_0) is now reachable, because (s_3, a_1) is influenced at $t = 1$ with 2-step influence. However, even though (s_1, a_0) now becomes influenced at $t = 1$, it cannot be reached by any influenced state-action pair, so (s_1, a_0) and (s_4, a_0) are influenced but unreachable.



(*a*) $k = 1$ (*b*) $k = 2$

Figure 3: Example counterfactual MDP given k-step influence. State-action pairs may or may not be influenced by the observed path, and states may or may not be reachable from other influenced state-action pairs.

We provide more details on the construction of the counterfactual MDPs in Figures 3(*a*) and 3(*b*) in Appendix D. We now re-formulate the idea of optimal counterfactual policy by incorporating our notion of influence. Previously, in Definition 1, we restricted to policies ensuring a bounded number of action changes from the observed action sequence. Here, we further include constraints to guarantee that any counterfactual path is influenced to some degree by the observation.

Definition 6 ((k, m)-CF policy) *Let τ be a path of an MDP $\mathcal{M}$ of length T, and let $\mathcal{M}^\tau$ be the corresponding counterfactual MDP. For a given $m \leq T - 1$ and influence bound k, a (k, m)-*CF policy *π^* is one that maximises the value $V_\tau(\pi) = \mathbb{E}_{\tau' \sim \mathcal{M}^\tau(\pi)} \left[\sum_{t=0}^{T-1} R(s'_t, a'_t) \right]$ under two conditions: 1) the observed path τ exerts a k-step influence on any counterfactual path τ' induced by $\mathcal{M}^\tau$ and π^*; and 2) any such counterfactual path satisfies the constraint $\sum_{t=0}^{T-1} \mathbb{1}(a_t \neq a'_t) \leq m$.*

By Remark 5, if $k \geq T + 1$, then a (k, m)-CF policy corresponds exactly to a m-CF policy, as defined in Definition 1, because the influence constraint will always be trivially satisfied ($t + k > T$, for all t).

4. Methodology

In this section, we describe the steps of our algorithm for finding the optimal (k, m)-CF policy for a given MDP. The pseudocode for the whole algorithm can be found in Appendix E. To achieve this, we first constrain the counterfactual MDP to only transitions that are influenced by the observed path (under a given k-step influence), then apply value iteration to find the optimal policy in the influence-constrained counterfactual MDP. Both these steps have polynomial complexity, and so the whole algorithm remains polynomial as well.

First, we calculate the counterfactual transition probabilities $P_{\mathcal{M},t,\tau}$ for all transitions using the top-down Gumbel sampling approach (Maddison et al., 2014), as described in Sections 2.1 and 2.2. To construct the influence-constrained counterfactual MDP, we first identify, for each state-action (s_t, a_t) pair in the observed path τ, all states s that are in the support of $P(\cdot \mid s_t, a_t)$. We denote the set of such states with S_t^τ, and denote the union of all of these S_t^τ as $S^\tau = \bigcup_{t=0}^{T-1} S_t^\tau$. For the example MDP given in Figures 3(*a*) and 3(*b*), $S^\tau = \{s_2, s_3, s_5, s_7\}$. Next, we execute a reverse Breadth-First Search (BFS) algorithm with a maximum depth of k over the original MDP, starting from each state in S^τ. This identifies the set $S^{\tau,k}$ of MDP states that can reach, within k steps, a state in S^τ, i.e., a state that is influenced by τ. As shown in Figure 3(*a*), $S^{\tau,1} = \{s_0, s_2, s_3, s_4, s_5, s_6, s_7\}$, and, as shown in Figure 3(*b*), $S^{\tau,2} = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$. Since the BFS algorithm is polynomial in the number of states and transitions, and we run the algorithm $|S^\tau| \leq |S|$ times, the worst-case computational complexity of the algorithm remains polynomial ($O(|S^\tau| \cdot (|S| + (|S|^2 \cdot |A|)))$).

To prune the non-influenced state-action pairs, we set $P_{\mathcal{M},t,\tau}(\cdot \mid s, a) = 0$ whenever there exists a s' where $P(s' \mid s, a) > 0$ in the original MDP, $s' \notin S^{\tau,k}$ and $t + k \leq T$ (as stated in Definition 4, all state-action pairs are influenced by the observation when $t + k > T$). This ensures that no action can lead to states outside the influence-constrained counterfactual MDP. Finally, we must prune any states that are unreachable or have no outgoing transitions, by setting $P_{\mathcal{M},t,\tau}(\cdot \mid s, a) = 0$ for all pairs (s, a) that have some probability of leading to these states. As shown in Figures 3(*a*) and 3(*b*), this last pruning step removes states s_4 and s_6 for $k = 1$, and states s_1 and s_4 for $k = 2$, as these states are not reachable.

Given the influence-constrained counterfactual MDP, we apply a value iteration algorithm (similar to the method in (Tsirtsis et al., 2021)) to find the optimal (k, m)-CF policy, with the only change being that the action choices are restricted to only those in the influence-constrained counterfactual MDP. The worst-case complexity of this dynamic programming algorithm for given values of k and m is $O(|\tau| \cdot |S|^2 \cdot |A| \cdot m)$, when all state-action pairs from the MDP $\mathcal{M}$ are contained in the influence-constrained counterfactual MDP. However, in practice, the size of the influence-constrained counterfactual MDP will often be significantly smaller than the unconstrained counterfactual MDP, reducing the complexity and execution time of the value iteration algorithm. The sizes of the counterfactual MDPs in our experiments, pruned with each value of k, are provided in Table 1.

Our approach is always guaranteed to identify the optimal (k, m)-CF policy, as follows:

Theorem 7 (Optimal (k, m)-CF Policy Guarantee) *For any given MDP $\mathcal{M}$, observed path τ, and values of k and m, our method is guaranteed to identify the optimal (k, m)-CF policy.*

Proof By construction, the pruning step of our algorithm ensures that only the parts of the counterfactual MDP that satisfy k-step influence remain. We then apply the value iteration algorithm from (Tsirtsis et al., 2021) to the influence-constrained counterfactual MDP to determine the optimal m-CF policy. This is a standard value iteration algorithm based on dynamic programming, which is

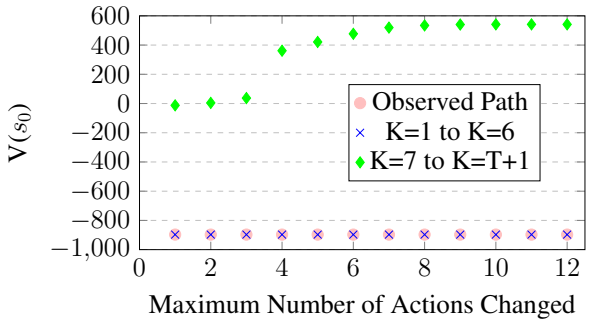


Figure 4: Grid World: value of initial state given k-step influence and maximum m actions changed

guaranteed to find the optimal policy (Sutton and Barto, 2018). Therefore, because the optimal m-CF policy is derived over only the transitions that satisfy the k-step influence constraint, our method is guaranteed to identify the optimal (k, m)-CF policy. ■

5. Experiments

In this section, we apply our notion of influence to several MDPs to derive counterfactual policies with varying levels of influence with respect to the observed path. We evaluate our approach on a Grid World model, an epidemic MDP model, and an MDP modelling sepsis patient trajectories[4].

5.1. Setup

For each MDP, we first generate an observed path of length T using a deliberately suboptimal policy. The choice of T depends on the MDP being evaluated to ensure there is sufficient opportunity to improve the policy, and is noted explicitly in the following subsections. We use this observed path to generate the optimal (k, m)-CF policy for $1 \leq k \leq T + 1$ and $1 \leq m \leq T$, using the algorithm described in Section 4. We assume that counterfactual paths begin in the same initial state as the observed path. To evaluate the performance of the (k, m)-CF policies, we consider the value of the initial state, $V(s_0)$, as this measures how good the paths generated by the optimal policy (starting at the initial state) will be. As a baseline, we consider the m-CF policies produced by the algorithm in (Tsirtsis et al., 2021), as this is the only other existing method for counterfactual inference in MDPs - this baseline is reported in our results as $k = T + 1$.

5.2. Grid World

In the Grid World experiment, the agent must traverse a 4x4 grid from the top-left corner to the bottom-right corner, avoiding a dangerous terminal state in the middle of the grid. At each time step, the agent can choose to move up, down, left, or right. However, there is a small probability that the agent will move in a different direction to the chosen action. As the agent gets closer to the goal state in the bottom-right corner, the reward for each state increases, and the agent receives a reward of 100

4. The source code for our experiments can be found at https://github.com/ddv-lab/counterfactual-influence-in-MDPs.

for reaching the terminal goal state. However, there is also a reward of -100 for transitioning into the dangerous terminal state.

We derive the counterfactual MDP using an observed path of length 11 that falls into this dangerous terminal state at $t = 3$. Figure 4 shows how k and m affect the value collected by the (k, m)-CF policy. When $k \leq 6$, we see no improvement in the policy, because all influenced counterfactual paths lead to the dangerous terminal state. However, when $k \geq 7$, the optimal influenced counterfactual path avoids the dangerous state for all m (shown by $V(s_0) \geq 0$), and when $m \geq 4$ it reaches the goal state and gains the $+100$ reward, resulting in much higher values for the initial state. This shows that we don't need to sacrifice optimality to generate counterfactual paths which are still influenced by the observed path, because the optimal counterfactual paths for $k = 7$ and $k = T + 1$ both reach the goal state, but the counterfactual path for $k = 7$ is more informed by the observation than the path for $k = T + 1$ (i.e., obtained without influence constraints).

5.3. Epidemic Model

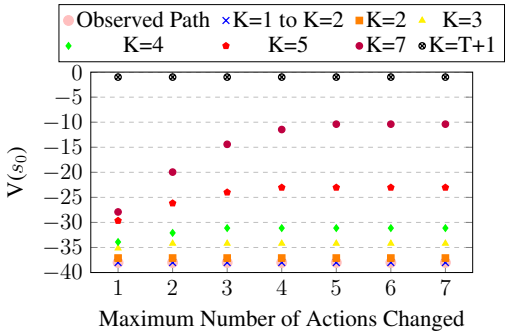
The epidemic MDP simulates how infection spreads through a (discrete) population. Each state (S, I, V) consists of S susceptible and I infected individuals, and the number of available vaccines V. At each time step, the agent implements a vaccination strategy and can choose among three actions: do nothing, vaccinate an infected individual, or vaccinate a susceptible individual. The reward for each transition (s, a, s') is given by the negative value of the number of infected individuals in s, $-I$. Full details of the model are given in Appendix F.

The observed path of length 7 begins in state $(S = 9, I = 1, V = 20)$ and is generated from a suboptimal policy that chooses to "do nothing" in every state. Therefore, we can generate increasingly better counterfactual paths that are still highly influenced by the observed path, because switching the action in most states would lead to an improved outcome. This is shown in Figure 5(*a*), where for a k-step influence with $k > 2$, the policy value increases monotonically with k and m. Figure 5(*b*) shows the average number of infected individuals I at each time step t for selected combinations of (k, m), compared to the observed path, for 1000 simulated trajectories. This further illustrates how relaxing the influence constraint impacts the final reward. With $k = 3$, we obtain a counterfactual path almost identical to the observed one except for the last two steps. On the other hand, when no influence constraints are imposed ($k = T + 1$), we obtain the optimal counterfactual path where the first action is changed to vaccinate the single infected person and stop the epidemic. Note that when the counterfactual paths follow the observed path (i.e., the counterfactual state coincides with the observed state, and the observed action is taken), the transitions become deterministic (leading to the next observed state with probability$= 1$), as we can see in the plot for some path prefixes.

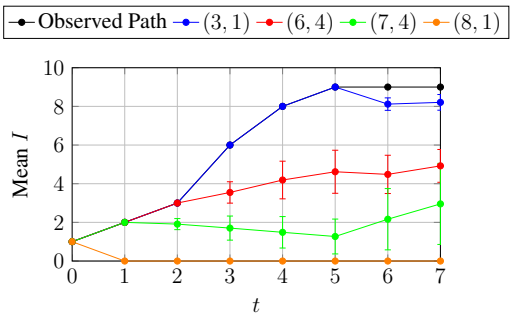
5.4. Sepsis Model

The Sepsis MDP is taken from (Oberst and Sontag, 2019)[5] and models the trajectories of sepsis patients. Each state consists of four vital signs (heart rate, blood pressure, oxygen concentration, and glucose levels), with possible values of low, normal, or high. There are three treatment options, which can be turned on or off at each time step (8 actions in total). Unlike in (Oberst and Sontag, 2019), we scale the rewards depending on the number of vital signs that are out of range, between -1000 (patient dies) and 1000 (patient is discharged). For further details, we refer to (Oberst and Sontag, 2019). In our experiments, we simulated the trajectory of a sepsis patient over 10 time steps.

5. Licensed under the MIT License, and available at https://github.com/clinicalml/gumbel-max-scm.

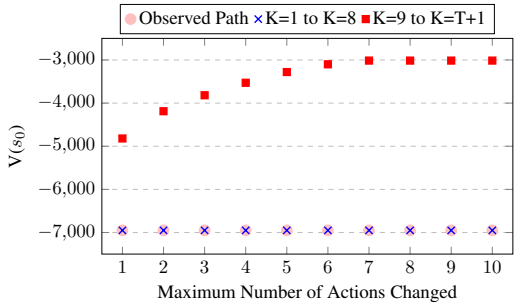


(*a*) Epidemic MDP: initial state value given changing influence

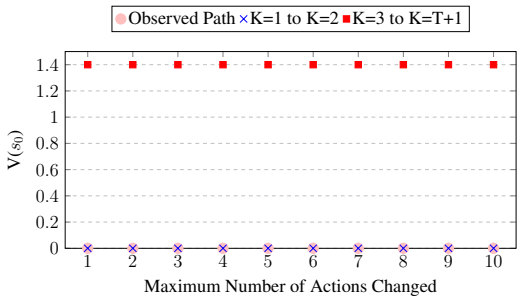


(*b*) Epidemic MDP: mean infection rate over time for select (k, m). The error bars show $\sigma(I)$.

Figure 5: Epidemic MDP analysis results



(*a*) Sepsis: $V(s_0)$ given k-step influence and maximum m actions changed for a catastrophic path.



(*b*) Sepsis: $V(s_0)$ given k-step influence and maximum m actions changed for a suboptimal path.

Figure 6: Sepsis MDP analysis results

In the model, the patient dies if three or more vital signs go out of range. This means that Sepsis trajectories depend heavily on the first few actions to reduce the probability that treatment will lead to one of these terminal states. However, if the observed path is a catastrophic one, as in Figure 6(*a*), changing these actions early on will often lead far away from the observed path, and so we need a low influence (i.e., high values of k) to obtain reasonably better outcomes. As shown in Figure 6(*b*), for suboptimal paths (where the patient is not dead/discharged at the end of the path, but is not discharged), a low value of k ($k \geq 3$) is sufficient to recover the optimal counterfactual path. However, only a small improvement on the observed path is possible.

5.5. Reduction in MDP Size

An additional benefit of our k-step influence approach is that the state space of the pruned counterfactual MDP can be significantly reduced. State space sizes for our models (for selected values of k) are given in Table 1. For $k = 1$, the MDP is restricted to just the states in the corresponding observed paths. We can also see that the state space of the pruned counterfactual MDPs for $k < T+1$ are significantly smaller (for the Epidemic and Sepsis models) than the state space of the entire

k	1	3	6	7	10	$T+1$	$\lvert S\rvert$
Grid World	12	15	15	147	182	192	16
Epidemic	8	43	157	210	-	19355	2541
Sepsis	11	14	3477	4055	5884	6996	1440

Table 1: Size of the state space of pruned counterfactual MDPs, given k-step influence. $|S|$ is the state space size of the original MDP. There is no data for the Epidemic environment for $k = 10$ because the observed path has length 7.

counterfactual MDP ($k = T + 1$), meaning that value iteration is much more efficient. Full results for all the environments are given in Appendix G, and execution times are discussed in Appendix H.

6. Conclusion

In this work, we addressed a significant yet neglected issue in counterfactual inference for Markov Decision Processes (MDPs): as counterfactual states and actions progressively diverge from the observed ones over time, the observation may no longer influence the counterfactual world, and the resulting explanation will no longer be tailored to the individual observation. To tackle this issue, we introduced a formal methodology to quantify the influence of the observed path τ on a counterfactual path τ', and devised an algorithm to generate optimal counterfactual explanations and policies while satisfying predetermined influence constraints. Our experiments reveal that while there exists a trade-off between influence and policy optimality, it is often possible to derive policies that are nearly optimal while still being significantly influenced by the initial observed path. The optimal degree of influence is domain-specific, but our method allows us to evaluate the trade-off between influence and optimality and make a better-informed choice on the value of the influence bound k. We include a detailed discussion on the choice of k in Appendix I.

Although this method is the first to expose and solve the problem of counterfactual influence, it relies on the availability of the system's transition probabilities. Moving forward, our goal is to develop optimal policies through a model-free approach, particularly in scenarios where the transition probabilities of the underlying MDP are unknown or uncertain.

Acknowledgments

This work was supported by UK Research and Innovation [grant number EP/W014785/2]; and UK Research and Innovation [grant number EP/S023356/1] in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org).

References

Casey C. Bennett and Kris Hauser. Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. *Artificial Intelligence in Medicine*, 57(1):9–19, 2013. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2012.12.003. URL https://www.sciencedirect.com/science/article/pii/S0933365712001510.

Nina L Corvelo Benz and Manuel Gomez Rodriguez. Counterfactual inference of second opinions. In *Uncertainty in Artificial Intelligence*, pages 453–463. PMLR, 2022.

Rita Borgo, Michael Cashmore, and Daniele Magazzeni. Towards providing explanations for AI planner decisions. In *IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI)*, 2018.

Lars Buesing, Theophane Weber, Yori Zwols, Nicolas Heess, Sebastien Racaniere, Arthur Guez, and Jean-Baptiste Lespiau. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *International Conference on Learning Representations*, 2018.

Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

Maria Fox, Derek Long, and Daniele Magazzeni. Explainable planning. *arXiv preprint arXiv:1709.10256*, 2017.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2052–2062. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Jasmina Gajcin and Ivana Dusparic. Acter: Diverse and actionable counterfactual sequences for explaining and diagnosing RL policies. *arXiv preprint arXiv:2402.06503*, 2024.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5084–5096. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jin21e.html.

Milad Kazemi, Jessica Lally, and Nicola Paoletti. Causal temporal reasoning for Markov decision processes. *Research Directions: Cyber-Physical Systems*, page 1–23, 2025. doi: 10.1017/cbp.2025.2.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21810–21823. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f7efa4f864ae9b88d43527f4b14f750f-Paper.pdf.

Taylor W Killian, Marzyeh Ghassemi, and Shalmali Joshi. Counterfactually guided policy transfer in clinical settings. In *Conference on Health, Inference, and Learning*, pages 5–31. PMLR, 2022.

Benjamin Krarup, Senka Krivic, Daniele Magazzeni, Derek Long, Michael Cashmore, and David E Smith. Contrastive explanations of plans through model restrictions. *Journal of Artificial Intelligence Research*, 72:533–612, 2021.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/b4288d9c0ec0a1841b3b3728321e7088-Paper.pdf.

Guy Lorberbom, Daniel D Johnson, Chris J Maddison, Daniel Tarlow, and Tamir Hazan. Learning generalized Gumbel-max causal mechanisms. *Advances in Neural Information Processing Systems*, 34:26792–26803, 2021.

Chaochao Lu, Biwei Huang, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, and Bernhard Schölkopf. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint arXiv:2012.09092*, 2020.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. *Advances in neural information processing systems*, 27, 2014.

Rahul Madhavan, Aurghya Maiti, Gaurav Sinha, and Siddharth Barman. Intervention efficient algorithm for two-stage causal MDPs. *arXiv preprint arXiv:2111.00886*, 2021.

Mausam Natarajan and Andrey Kolobov. *Planning with Markov decision processes: An AI perspective*. Springer Nature, 2022.

Kimia Noorbakhsh and Manuel Rodriguez. Counterfactual temporal point processes. *Advances in Neural Information Processing Systems*, 35:24810–24823, 2022.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

Jonathan Patrick and Mehmet A Begen. Markov decision processes and its applications in healthcare. *Handbook of healthcare delivery systems. CRC, Boca Raton*, 2011.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Mark Shifrin and Hava Siegelmann. Near-optimal insulin treatment for diabetes patients: A machine learning approach. *Artificial Intelligence in Medicine*, 107:101917, 2020. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2020.101917. URL https://www.sciencedirect.com/science/article/pii/S0933365719305640.

Gregory Stein. Generating high-quality explanations for navigation in partially-revealed environments. *Advances in neural information processing systems*, 34:17493–17506, 2021.

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

Stelios Triantafyllou, Aleksa Sukovic, Debmalya Mandal, and Goran Radanovic. Agent-specific effects: A causal effect propagation analysis in multi-agent MDPs. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=pmncWWkGMz.

Stratis Tsirtsis and Manuel Rodriguez. Finding counterfactually optimal action sequences in continuous state spaces. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 3220–3247. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/09ae6beae5f1ff38f05c05979097ea0f-Paper-Conference.pdf.

Stratis Tsirtsis, Abir De, and Manuel Rodriguez. Counterfactual explanations in sequential decision making under uncertainty. *Advances in Neural Information Processing Systems*, 34:30127–30139, 2021.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. A survey on causal reinforcement learning. *arXiv preprint arXiv:2302.05209*, 2023.

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8252831b9fce7a49421e622c14ce0f65-Paper.pdf.

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

Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. Counterfactual off-policy training for neural dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3438–3448, 2020.

Appendix A. Simulation Evidence for Motivating Example

In Figure 1, we showed that if we allow counterfactual paths to be unconstrained, these paths may diverge from the observation such that their counterfactual transition probabilities are uninformed by the observation. Consequently, these counterfactual paths will no longer be tailored to the specific observation. Deriving optimal counterfactual policies over areas of the counterfactual MDP that are not influenced by the observation could yield policies that are not tailored to the given observation, and may actually be suboptimal for the particular observation.

This is particularly an issue when hidden subgroup differences exist within the population. In learned MDPs, some aspects of the true, underlying state could remain unobserved, which can lead to differences in the transitions taken by different subgroups. Therefore, in these environments, it is crucial that counterfactual policies are tailored to the observation, or they may be optimal for the population as a whole, but suboptimal for the particular subgroup that the observation belongs to.

We can evaluate this potential suboptimality by considering an example MDP where we have access to the transition probability and reward functions for both a fully observable and partially observable version of the MDP. This partially observable version of the MDP represents an MDP that we may be able to learn in practice (where, for example, we can only observe an incomplete set of variables in the learned MDP). Given an observed trajectory, we can learn:

1. The optimal counterfactual policy across the general population (i.e., the unconstrained counterfactual policy), using the partially observable MDP.
2. Various k-CF policies, again using the counterfactual partially observable MDP.
3. The “true" optimal counterfactual policy for the diabetic patient, using the fully observable MDP.

We can then compare these policies by measuring the average cumulative reward they achieve over the “true" counterfactual MDP (i.e., the fully observable counterfactual MDP). We expect the k-CF policies (2) to approximate the subgroup-specific policy (3) more closely than the general-population policy (1) (and therefore achieve higher average rewards than the general-population policy) because these policies will be restricted to areas of the counterfactual MDP that are more informed/influenced by the observation.

This effect is particularly noticeable where the observation is optimal or close to optimal. This is because these observed paths typically require very few (or no) changes to improve upon the observation, hence the optimal counterfactual paths will be close to the observation in the counterfactual MDP. When we generate counterfactual policies with a small value of k, this will restrict the counterfactual MDP to those areas that are highly tailored to the observation, and will include these higher reward counterfactual paths.

As an example, we take the sepsis example from Figure 1. In the Sepsis MDP (Oberst and Sontag, 2019), the diabetic status of the patient can be explicit or hidden in the state. Given an observed trajectory of a diabetic patient, if we derive the optimal counterfactual policy for this observation over the partially observable MDP, without constraining the policy to areas of the counterfactual MDP that are sufficiently influenced by the observation, this may lead to policies that are optimal for the general population and not optimal for the observed diabetic patient.

Figures 7 and 8 present the average cumulative reward obtained by these policies on two observed diabetic trajectories. These trajectories are of length $T = 10$, hence $k = 11$ corresponds to the entire

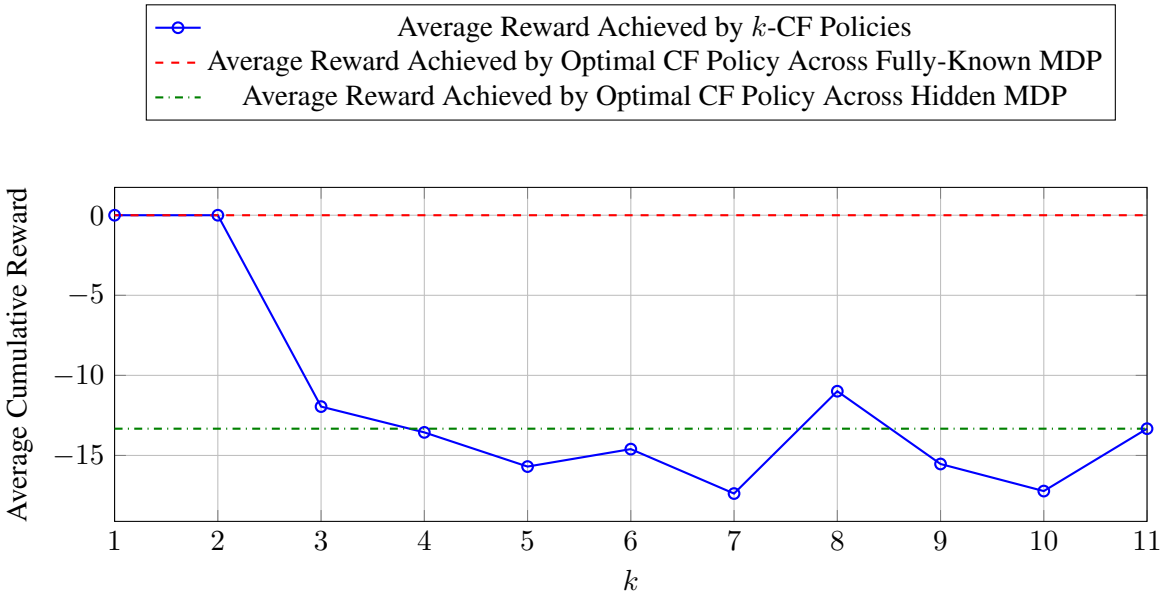


Figure 7: Average cumulative reward of policies, given an observed diabetic path under the optimal policy.

counterfactual MDP (see Section 3). Figure 7 compares the policies on an observed diabetic path under the optimal policy. As expected, the optimal counterfactual policy over the fully observable MDP is the same as the observed policy, as are the $k = 1$ and $k = 2$ policies, which are derived from areas of the counterfactual MDP that are greatly influenced by the observation. However, the average cumulative reward achieved by the optimal counterfactual policy across the partially observable MDP is much lower, as this is optimal across the general population rather than for diabetic patients. We also see a decline in performance for k-CF policies where $k \geq 3$, as these policies are derived over areas of the counterfactual MDP that are less influenced by (and therefore less tailored to) the observation.

Figure 8 compares the policies on an observed diabetic path under a suboptimal policy (the optimal policy with some randomly chosen actions). As expected, we see that the average cumulative reward achieved by the optimal counterfactual policy over the fully observable MDP is again higher than that of the partially observable MDP. We also see that there is a decline in performance for counterfactual policies with higher values of k (in this case $k \geq 10$), as these are learnt over areas of the counterfactual MDP that are not very influenced by the observation, and therefore are closer to the general population counterfactual policy rather than the “true" diabetic counterfactual policy.

Appendix B. Related Work

To the best of our knowledge, there is no other work that directly aims to address the problem of counterfactual influence caused by the divergence of the counterfactual path from the observed one. Nevertheless, there is growing field of work focusing on the intersection between causality and various other domains, including reinforcement learning and planning.

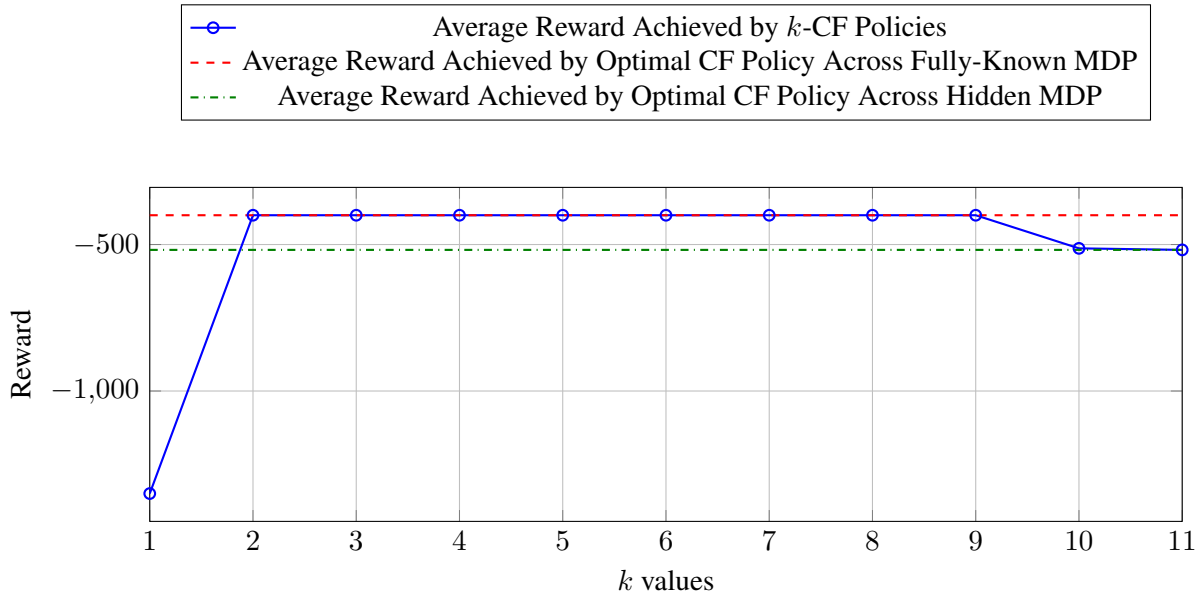


Figure 8: Average cumulative reward of policies, given an observed diabetic path under a suboptimal policy.

Causal RL Causality can often improve the performance of RL algorithms (Zeng et al., 2023; Schölkopf et al., 2021), especially when data is scarce, or where exploration may be dangerous or infeasible (Lu et al., 2020). In such cases, counterfactual reasoning can be used to augment datasets with counterfactual data, improving the efficiency and performance of RL algorithms (Lu et al., 2020; Buesing et al., 2018), to generate counterfactual paths as a causal explanation for how an observed policy could be improved (Oberst and Sontag, 2019; Tsirtsis et al., 2021; Tsirtsis and Rodriguez, 2023; Gajcin and Dusparic, 2024), or to measure the influence of an individual agent's action/treatment decision on the outcome in a multi-agent setting (Triantafyllou et al., 2024) (a different notion of influence to the notion of counterfactual influence in this paper).

Causal reasoning is also useful at training-time: if an agent can perform informative interventions to learn the causal structure of its environment, it would enable performing more structured exploration when learning optimal policies (Dasgupta et al., 2019). In addition to MDPs, causal RL has also been successfully applied to Multi-Arm Bandit problems (Lattimore et al., 2016; Madhavan et al., 2021) and Dynamic Treatment Regimes (Zhang and Bareinboim, 2019).

Planning Our work is related to the field of explainable planning (Fox et al., 2017) and in particular contrastive explanations (Borgo et al., 2018), which focuses on offering explanations about alternative sequence of actions in classic planning scenarios. Krarup et al. (2021) propose a method to restrict the model by implementing constraints based on user questions, thereby providing structured explanations for the planning procedure as a negotiation. Stein (2021) extends this to explanations for plans in partially-revealed environments. However, it should be noted that these counterfactual explanations do not adjust the transition probabilities based on the observed path in the counterfactual world.

Offline RL In offline RL, the objective is to find an optimal policy maximising the expected return using a fixed dataset of observed trajectories (Uehara et al., 2022). However, this can be challenging due to the problem of *distribution shift*, where there is a mismatch between the distribution of trajectories in the dataset, and distribution of the trajectories that would be generated by the learned policy (Jin et al., 2021). This often leads to overestimation of the value function for out-of-distribution actions (Uehara et al., 2022). Recent work tackles distribution shift by promoting proximity between the learned policy and the behaviour policy. This is achieved through regularising the learned policy to avoid states and actions that appear less frequently (Fujimoto et al., 2019), or using pessimistic value-based approaches, which apply a penalty to the value function on these states and actions (Yu et al., 2020; Kidambi et al., 2020). Although our work solves a different problem, our notion of influence is similar to these methods as it can be seen as form of regularisation constraint.

Appendix C. Discussion on Notion of Influence

In our work, we formulated our notion of influence from a structural perspective, in terms of the supports of the state-action pairs. By Proposition 2, we know this is an efficient and precise condition for influence. However, one could argue that this notion is too restrictive as this exerts hard constraints for pruning. Alternatively, we could have formulated 1-step influence from a probabilistic perspective. But, any notion would have to be in terms of the interventional probability distributions alone: any notion using the counterfactual probabilities (e.g., the statistical distance between the nominal and counterfactual distributions) would not be exact, due to sampling variability in the counterfactual probabilities from the sampled Gumbel values. For example, we could measure 1-step influence as:

$$\frac{\sum_{s' \in \mathcal{S}} |P(s' \mid s_t, a_t) - P(s' \mid s, a)|}{2} = \begin{cases} 1 \text{ if the distributions have disjoint support} \\ < 1 \text{ if the distributions have overlapping support} \end{cases}$$

and specify some ϵ as the maximum of this sum. Similar to our notion of k-step influence, we may want to extend this to paths, e.g., ensuring the total statistical distance is less than some value. We could also consider the reward: how should we find an appropriate ϵ that balances a path's total statistical distance with the total reward of that path, e.g., how should we choose parameters α and β below:

$$\alpha \frac{\sum_{s' \in \mathcal{S}} |P(s' \mid s_t, a_t) - P(s' \mid s, a)|}{2} - \beta \sum_{s' \in \mathcal{S}} r(s, a, s') - r(s_t, a_t, s')$$

This is a design choice that depends on how safety-critical the domain is: ensuring the counterfactual paths are sufficiently informed vs. optimising the total reward. We chose to use our structural notion in this paper, as it is a simple and exact notion to define influence. But, in future work, it would be interesting to compare these two notions of influence, to see in what situations these two notions differ.

One simple MDP example that illustrates the differences between these notions is given in Figure 9. The observed path is given in red, and the transitions that would be contained in the influence-constrained counterfactual MDP (under our structural notion of influence for $k = 2$) are represented by the solid arrows. The influence-constrained counterfactual MDP is quite restrictive, largely due to the transition from $s_3 \rightarrow s_7$ which deviates far from the observed path. Under a probabilistic notion

of influence, we might instead choose to ignore that this transition deviates far from the observed path because it has such low probability ($p = 0.01$), and consider all of the counterfactual MDP. However, if the nominal probability of the transition from $s_3 \to s_7$ was much higher (e.g., $p > 0.1$), we may want the probabilistic notion of influence to remove this path, unless, for example, the reward for reaching state s_{11} was very high. The trade-off between influence and reward is a design choice and is domain-dependent, so any notion of influence should consider this: our structural notion of influence has the influence bound k that can be changed to loosen the restriction on influence, and achieve higher rewards.

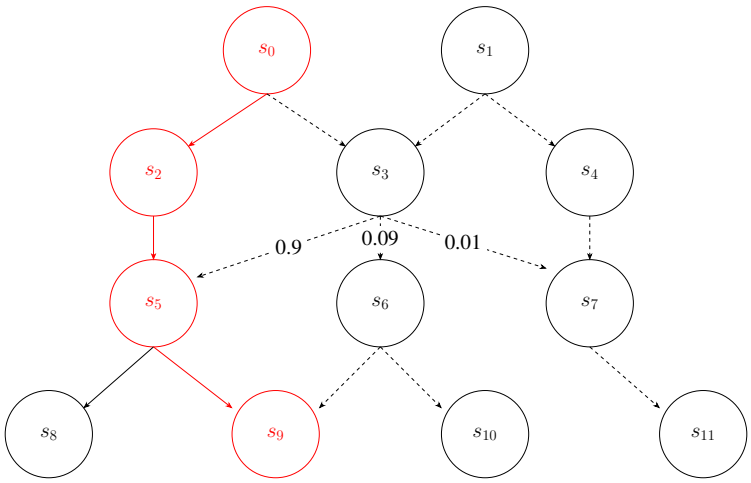


Figure 9: Simple MDP example to illustrate differences between structural and probabilistic notions of influence. The observed path is given in red.

Appendix D. Example of Influence-Constrained Counterfactual MDP

Take the counterfactual MDP example from Figures 3(*a*) and 3(*b*). This has two possible actions, a_0 and a_1, at states s_0 and s_3, and only action a_0 at the rest of the states. The full transition table for the nominal MDP is given in Table 2. The support of each state-action pair is the set of states that can be reached (with non-zero probability) from the state-action pair, and is given in Table 3.

State	Action	Next State	Transition Probability
s_0	a_0	s_2	0.5
s_0	a_0	s_3	0.5
s_0	a_1	s_1	1.0
s_1	a_0	s_4	1.0
s_2	a_0	s_5	1.0
s_3	a_0	s_5	1.0
s_3	a_1	s_6	1.0
s_4	a_0	s_7	1.0
s_5	a_0	s_7	1.0
s_6	a_0	s_7	1.0

Table 2: Nominal MDP transition table

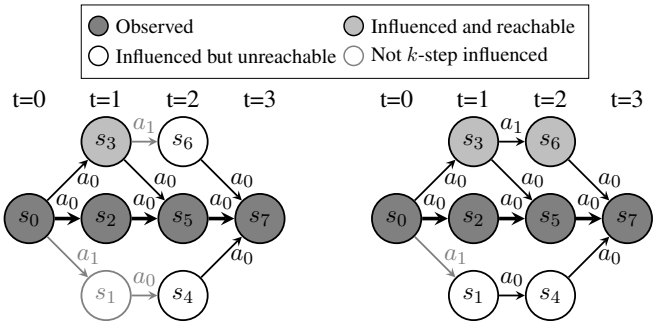
State	Action	Support
s_0	a_0	$\{s_2, s_3\}$
s_0	a_0	$\{s_3\}$
s_0	a_1	$\{s_1\}$
s_1	a_0	$\{s_4\}$
s_2	a_0	$\{s_5\}$
s_3	a_0	$\{s_5\}$
s_3	a_1	$\{s_6\}$
s_4	a_0	$\{s_7\}$
s_5	a_0	$\{s_7\}$
s_6	a_0	$\{s_7\}$

Table 3: Supports of state-action pairs in the nominal MDP

Given the observed path $\tau = [(s_0, a_0), (s_2, a_0), (s_5, a_0), (s_7, a_0)]$, we can now find the influence-constrained counterfactual MDP given $k = 1$ and $k = 2$. When $k = 1$ (Figure 10(*a*)), (s_1, a_0) and (s_3, a_1) are not influenced at $t = 1$, because they have disjoint supports ($\{s_4\}$ and $\{s_6\}$ respectively) with the observed state-action pair (s_2, a_0) (whose support is $\{s_5\}$). For the opposite reason, (s_2, a_0) and (s_3, a_0) are influenced at $t = 1$, as the supports of (s_2, a_0) and (s_3, a_0) are both $\{s_5\}$.

(s_4, a_0), (s_5, a_0) and (s_6, a_0) are all influenced at $t = 2$, as their supports overlap with the observed pair (s_5, a_0) (in fact, all of their supports are exactly $\{s_7\}$). However, even though (s_4, a_0) and (s_6, a_0) are influenced, they cannot be reached from any influenced state-action pairs, so are also removed from the influence-constrained counterfactual MDP: we say they are "influenced but unreachable".

Figure 10(*b*) depicts the case of 2-step influence. We note that (s_6, a_0) is now reachable, because although the support of (s_3, a_1) is disjoint from the support of the observed state-action pair (s_2, a_0) (and so is not 1-step influenced), all transitions leading to (s_3, a_1) and from the states that (s_3, a_1) can reach are influenced, so (s_3, a_1) is influenced at $t = 1$ with 2-step influence. However, even though (s_1, a_0) now becomes influenced at $t = 1$, it cannot be reached by any influenced state-action pair, so (s_1, a_0) and (s_4, a_0) are influenced but unreachable.



(*a*) $k = 1$ (*b*) $k = 2$

Figure 10: Example counterfactual MDP given k-step influence. State-action pairs may or may not be influenced by the observed path, and states may or may not be reachable from other influenced state-action pairs.

Appendix E. Algorithm for Constructing Influence-Constrained Counterfactual MDP

Algorithm 1 Find Optimal (k, m)-CF Policy for a given MDP

Input: MDP transition probabilities P, observed path τ, counterfactual transition probabilities $P_{\mathcal{P},t,\tau}$, k, m
Output: Optimal (k, m)-CF policy π^*
$S^\tau \leftarrow \emptyset$ {Initialise set of all states in support of each observed (s_t, a_t)}
for each state-action pair (s_t, a_t) in the observed path **do**
  $S_t^\tau \leftarrow$ all states in the support of $P(\cdot \mid s_t, a_t)$: $\{s' \mid P(s' \mid s_t, a_t) > 0\}$
  $S^\tau \leftarrow S^\tau \cup S_t^\tau$
end for
$S^{\tau,k} \leftarrow \emptyset$ {Initialise set of all states which are k-step influenced}
for each state s in S^τ **do**
  $S^{\tau,k} \leftarrow S^{\tau,k} \cup \text{ReverseBFS}(s, k)$ {Reverse BFS with depth k}
end for
for t in range$(0, T - k + 1)$ **do**
  for each s, a, s' **do**
    if $P(s' \mid s, a) > 0$ and $s' \notin S^{\tau,k}$ **then**
      $P_{\mathcal{P},t,\tau}(\cdot \mid s, a) = 0$ {Prune non-influenced transitions}
    end if
  end for
end for
Further prune MDP to remove transitions leading to unreachable states and states with no outgoing edges
Compute the optimal (k, m)-CF policy using dynamic programming, while restricting action choices to transitions in the influence-constrained counterfactual MDP
return {Optimal (k, m)-CF policy}

Appendix F. Epidemic MDP

The Epidemic MDP models how infection spreads through a given population P of vaccinated and unvaccinated individuals. The MDP uses a hypergeometric distribution to model how many susceptible individuals become infected at each step. Each vaccination decreases the count V and removes the vaccinated individual from the population (i.e., no re-infection is possible). The reward for each transition (s, a, s') is given by the negative of the number of infected individuals in s, $-I$.

The MDP can be described as follows.

State Space The state space consists of a tuple (S, I, V) where:

- S: number of individuals susceptible to the disease (ranging from 0 to P).
- I: number of individuals infected with the disease (ranging from 0 to P).
- V: number of vaccines available (ranging from 0 to $2 \times P$).

Initial State The initial state (S_0, I_0, V_0) consists of:

- $S_0 = P - I_0$ (initially the entire population is unvaccinated).
- I_0 is chosen arbitrarily or can be taken from any chosen distribution. In our experiments, we set $I_0 = 1$.
- $V_0 = 2 \times P$.

Action Space There are three possible actions at each time step:

- V_I: vaccinate an infected individual.
- V_S: vaccinate a susceptible individual.
- Nil: do nothing.

Transition Probabilities The transition probabilities are defined as follows. We assume that at each time step, individuals in S_t can be infected following a hypergeometric model, i.e., a binomial without replacement.

- For the action Nil:
  - $P(S_{t+1}, I_{t+1}, V_{t+1} \mid S_t, I_t, V_t, NIL) = 0$ if $V_{t+1} \neq V_t$.
  - For $k \leq S_t$, $P(S_{t+1}{-}k, I_{t+1}{+}k, V_{t+1} \mid S_t, I_t, V_t, \text{NIL}) = \text{hypergeom}(M, n, N).\text{pmf}(k)$ if $V_{t+1} = V_t$, where $M = S_t + I_t$, $n = \min(S_t, I_t)$, $N = S_t$.
- For the action V_I:
  - $P(S_{t+1}, I_{t+1}, V_{t+1} \mid S_t, I_t, V_t, V_I) = 0$ if $V_{t+1} \neq V_t - 1$.
  - For $k \leq S_t$, $P(S_{t+1}{-}k, I_{t+1}{-}1{+}k, V_{t+1} \mid S_t, I_t, V_t, V_I) = \text{hypergeom}(M, n, N).\text{pmf}(k)$ if $V_{t+1} = V_t - 1$, where $M = S_t + I_t - 1$, $n = \min(S_t, I_t - 1)$, $N = S_t$.
- For the action V_S:
  - $P(S_{t+1}, I_{t+1}, V_{t+1} \mid S_t, I_t, V_t, V_S) = 0$ if $V_{t+1} \neq V_t - 1$.
  - For $k \leq S_t{-}1$, $P(S_{t+1}{-}k{-}1, I_{t+1}{+}k, V_{t+1} \mid S_t, I_t, V_t, V_S) = \text{hypergeom}(M, n, N).\text{pmf}(k)$ if $V_{t+1} = V_t - 1$, where $M = S_t + I_t - 1$, $n = \min(S_t - 1, I_t)$, $N = S_t - 1$.

Rewards The reward function at each time step t is defined as the negative of the number of infected individuals, $R_t = -I_t$.

Appendix G. Size of State Space of Pruned Counterfactual MDPs, Given *k*-step Influence

Table 4: Grid World: Size of the State Space of Pruned Counterfactual MDP, Given k-step influence

k	1	2	3	4	5	6	7	8	9	10	11	T+1	|S|
State Space	12	12	15	15	15	15	147	161	173	182	188	192	16

Table 5: Epidemic: Size of the State Space of Pruned Counterfactual MDP, Given k-step influence

k	1	2	3	4	5	6	7	T+1	|S|
State Space	8	32	43	59	91	157	210	19355	2541

Table 6: Sepsis: Size of the State Space of Pruned Counterfactual MDP, for the Catastrophic Path, Given k-step influence

k	1	2	3	4	5	6	7	8	9	10	T+1	|S|
State Space	11	14	14	2123	2896	3477	4055	4647	5291	5884	6996	1440

Appendix H. Training Details

Our algorithm was implemented in Python 3.10 and executed on a 128-core machine with an Intel Xeon CPU and 512 GB RAM, but only 32 threads were required to calculate the counterfactual transition probabilities, which was the only parallelised part of the algorithm.

The Grid World case study was relatively quick as this has a relatively small state space: for a fixed choice of (k, m), deriving the (pruned) counterfactual MDP and computing the optimal policy runs in the order of minutes. The Epidemic and Sepsis case studies have larger state spaces, and so it took several hours to derive the counterfactual MDP and run policy iteration for every combination of (k, m).

Appendix I. Discussion on Choice of k

The parameter k sets the level of influence that we consider 'sufficient' for our counterfactual paths to be informed by the observation. The choice of k is domain-dependent, and there may not necessarily be a "correct" value of k. Instead, we consider k to be a design choice. For example, in healthcare and other safety-critical domains, it is desirable that any counterfactual path (which will be used as a counterfactual explanation for how the current treatment policy could be improved) is well informed by the observation: this would naturally lead to choosing low k values. However, for less safety-critical domains, we are more concerned about optimising the reward, at the risk of doing so over non-influenced paths (i.e., paths that are not tailored to the observation). In such cases, we may want to choose higher k values, (e.g., the smallest k s.t. the counterfactual reward meets some threshold).

In some domains, it may be possible to select an appropriate k for a particular observation. For example, in the Sepsis MDP experiment, you can identify counterfactual policies and counterfactual paths for different values of k, and a domain expert (e.g., a clinician) could assess these paths and identify whether they would be realistic or unrealistic for the observed patient, thereby allowing us to identify counterfactual policies that are tailored to the individual. For example, in our Sepsis example in Figure 1, a clinician might be able to tell that the unconstrained counterfactual path is unrealistic for the observed trajectory of the diabetic patient, e.g., because the patient's blood sugar levels in the unconstrained counterfactual path look "too stable" for a diabetic patient with Sepsis. By comparing counterfactual trajectories generated at different levels of k, the clinician may be able to set a maximum k below which the counterfactual paths appear reliable, based on their expert knowledge and experience.

However, in other domains, it may be more challenging to evaluate whether a counterfactual path remains tailored to the particular observation. Here, the choice of k will depend on the given task. If adherence to the observation is important (e.g., personalised therapy), a small value of k may be preferable to ensure that any policy changes remain informed by the original observation. On the other hand, in less safety-critical domains, we may choose a large value of k to allow for larger deviations from the observation, particularly if the observed path was catastrophic, in the hopes of achieving higher rewards.