

## 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

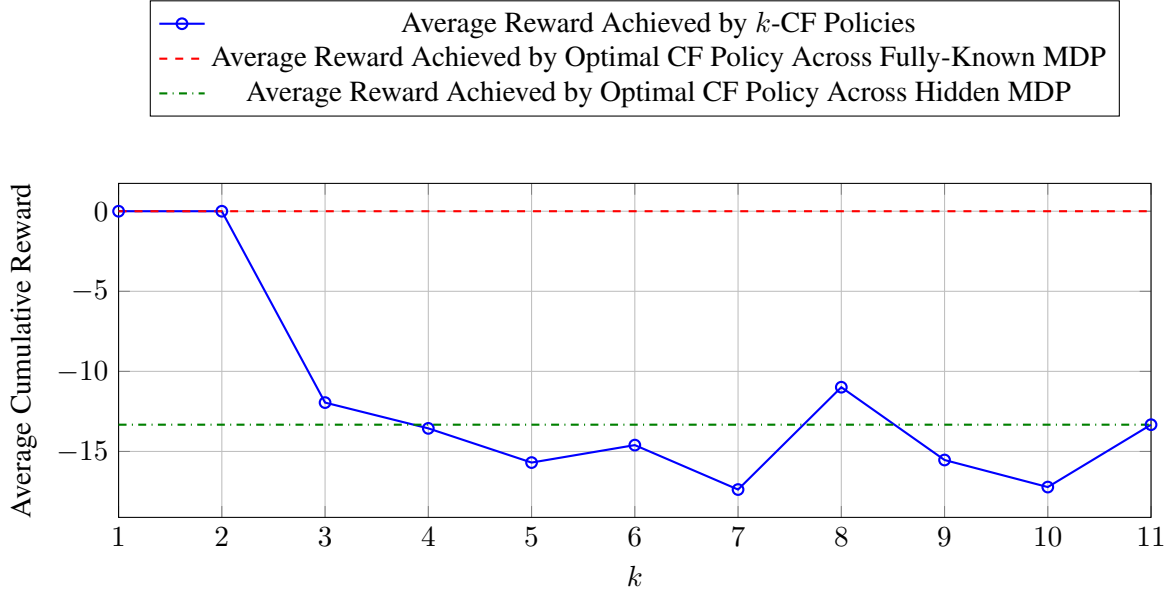


Figure 7: Average cumulative reward of policies, given an observed diabetic path under the optimal policy.

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.

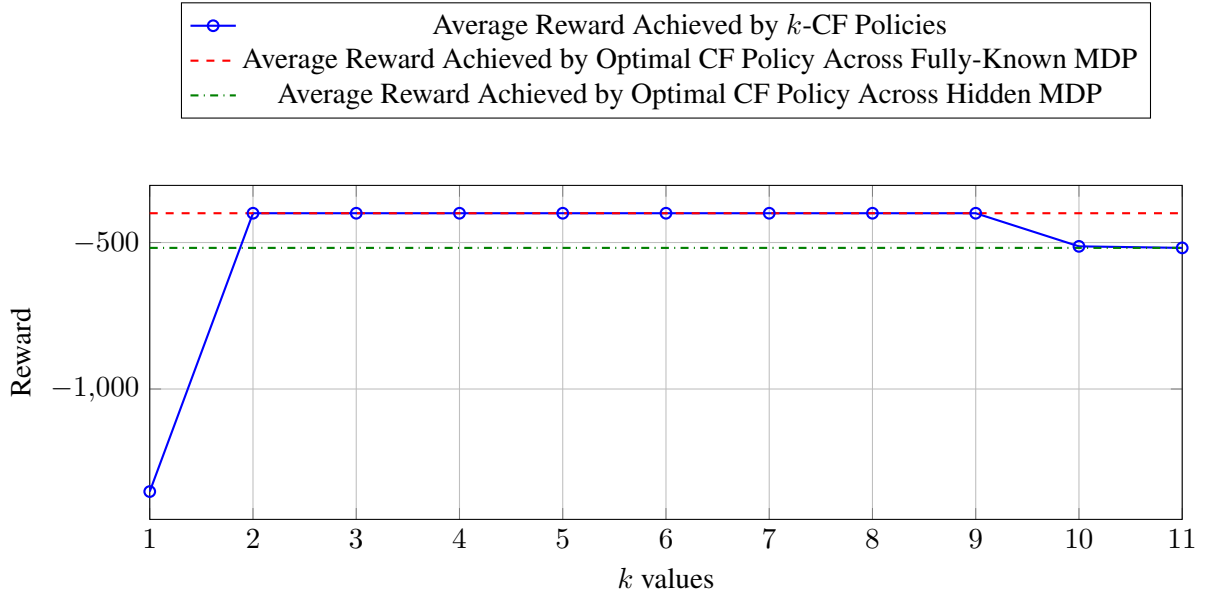


Figure 8: Average cumulative reward of policies, given an observed diabetic path under a suboptimal policy.

**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:

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

and specify some  $\epsilon$  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  $\epsilon$  that balances a path’s total statistical distance with the total reward of that path, e.g., how should we choose parameters  $\alpha$  and  $\beta$  below:

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

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  $s_3 \rightarrow s_7$  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.

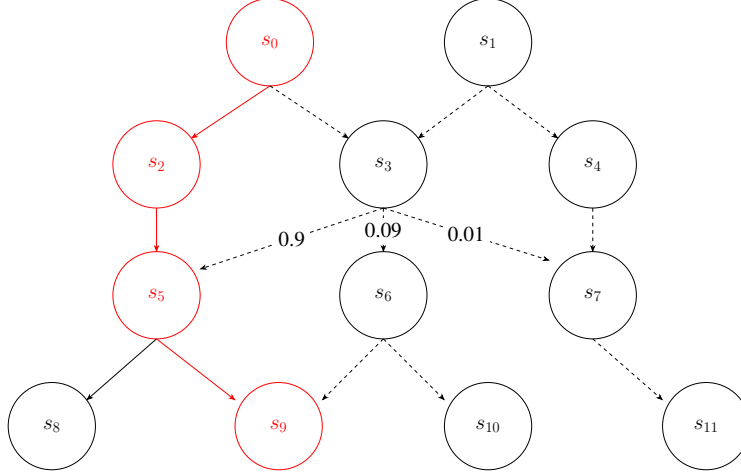


Figure 9: Simple MDP example to illustrate differences between structural and probabilistic notions of influence. The observed path is given in red.

#### 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
$s_0$	$a_0$	$\{s_2, s_3\}$
$s_0$	$a_1$	$\{s_1\}$
$s_1$	$a_0$	$\{s_4\}$
$s_2$	$a_0$	$\{s_5\}$
$s_3$	$a_0$	$\{s_5\}$
$s_3$	$a_1$	$\{s_6\}$
$s_4$	$a_0$	$\{s_7\}$
$s_5$	$a_0$	$\{s_7\}$
$s_6$	$a_0$	$\{s_7\}$

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.

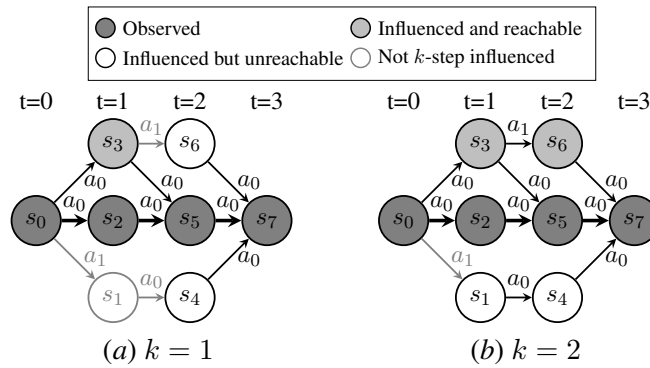


Figure 10: Example counterfactual MDP given  $k$ -step influence. State-action pairs may or may not be influenced by the observed path, and states may or may not be reachable from other influenced state-action pairs.

## Appendix E. Algorithm for Constructing Influence-Constrained Counterfactual MDP

---

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

---

```

1: Input: MDP transition probabilities  $P$ , observed path  $\tau$ , counterfactual transition probabilities  $P_{\mathcal{P},t,\tau}$ ,  $k, m$ 
2: Output: Optimal  $(k, m)$ -CF policy  $\pi^*$ 
3:  $S^\tau \leftarrow \emptyset$  {Initialise set of all states in support of each observed  $(s_t, a_t)$ }
4: for each state-action pair  $(s_t, a_t)$  in the observed path do
5:    $S_t^\tau \leftarrow$  all states in the support of  $P(\cdot \mid s_t, a_t)$ :  $\{s' \mid P(s' \mid s_t, a_t) > 0\}$ 
6:    $S^\tau \leftarrow S^\tau \cup S_t^\tau$ 
7: end for
8:  $S^{\tau,k} \leftarrow \emptyset$  {Initialise set of all states which are  $k$ -step influenced}
9: for each state  $s$  in  $S^\tau$  do
10:   $S^{\tau,k} \leftarrow S^{\tau,k} \cup \text{ReverseBFS}(s, k)$  {Reverse BFS with depth  $k$ }
11: end for
12: for  $t$  in range( $0, T - k + 1$ ) do
13:   for each  $s, a, s'$  do
14:    if  $P(s' \mid s, a) > 0$  and  $s' \notin S^{\tau,k}$  then
15:       $P_{\mathcal{P},t,\tau}(\cdot \mid s, a) = 0$  {Prune non-influenced transitions}
16:    end if
17:   end for
18: end for
19: Further prune MDP to remove transitions leading to unreachable states and states with no outgoing edges
20: Compute the optimal  $(k, m)$ -CF policy using dynamic programming, while restricting action choices to transitions in the influence-constrained counterfactual MDP
21: 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  $k \leq S_t$ ,  $P(S_{t+1}-k, I_{t+1}+k, V_{t+1} \mid S_t, I_t, V_t, NIL) = \text{hypergeom}(M, n, N).pmf(k)$  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  $k \leq S_t$ ,  $P(S_{t+1}-k, I_{t+1}-1+k, V_{t+1} \mid S_t, I_t, V_t, V_I) = \text{hypergeom}(M, n, N).pmf(k)$  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  $k \leq S_t - 1$ ,  $P(S_{t+1}-k-1, I_{t+1}+k, V_{t+1} \mid S_t, I_t, V_t, V_S) = \text{hypergeom}(M, n, N).pmf(k)$  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.