All's Well That Ends Well: Avoiding Side Effects with Distance-Impact Penalties

Charlie Griffin* **Joar Skalse** **Lewis Hammond** **Alessandro Abate**

Department of Computer Science
University of Oxford

cg@charlie.griffin.me, {joar.skalse,lewis.hammond,aabate}@cs.ox.ac.uk

Abstract

Misspecifying the reward function of a reinforcement learning agent may cause catastrophic side effects. In this work, we investigate *distance-impact penalties*: a general-purpose auxiliary reward based on a state-distance measure that captures, and thus can be used to penalise, side effects. We prove that the size of the penalty depends only on an agent's final impact on the environment. Distance-impact penalties are scalable, general, and immediately compatible with model-free algorithms. We analyse the sensitivity of an agent's behaviour to the choice of penalty, expanding results about reward-shaping, proving sufficient and necessary conditions for policy-optimality to be invariant to misspecification, and providing error bounds for optimal policies. Finally, we empirically investigate distance-impact penalties in a range of grid-world environments, demonstrating their ability to prevent side effects whilst permitting task completion.

1 Introduction

Reinforcement learning (RL) is a general framework for sequential decision-making that may produce super-human performance when there is a clear objective (Silver et al., 2016). However, even in simple environments, it can be hard to design a reward function that captures our true objective (Hendrycks et al., 2022). Optimising a proxy objective can lead to unpredictable or dangerous behaviour and unintended *side effects* (Krakovna et al., 2020a). Since it is difficult to anticipate and penalise each possible side effect in advance, Armstrong and Levinstein (2017) suggest using *impact regularisers*: general training methods that bias an agent towards changing the environment as little as possible.

There are many challenges to choosing an appropriate impact regulariser (Amodei et al., 2016). An overpowered regulariser may prevent the agent from completing its task but an underpowered one may not be sufficient to prevent side effects. Further, Krakovna et al. (2020b) showed that some poorly designed impact measures cause agents to obstruct humans (e.g. by preventing them from changing the environment by eating food in it). It is therefore important to investigate how and when impact regularisers are effective.

In this paper, we investigate distance-impact penalties: a way to penalise an agent's actions by exactly their ultimate effect on the environment. Unlike previous methods (Krakovna et al., 2020b; Turner et al., 2020), distance-impact penalties *reward* behaviour that reverses previous side effects and reduce the task of augmenting a goal-oriented reward function with safety considerations to the (arguably easier) task of designing a distance measure that characterises the magnitude of difference between

*Corresponding author.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

any two world-states. Our work differs from prior use of potential-based regularisers (Vamplew et al., 2021) in its analysis of when, and by how much, misspecifying the distance measure affects the optimal policy. We empirically demonstrate that distance-impact penalties can create low-impact agents with deep Q-learning (Mnih et al., 2015), and investigate the robustness of learnt policies' behaviours to distance-measure design. Finally, we compare distance-impact measures to existing methods and suggest directions for future work.

1.1 Definitions

We characterise RL problems using Markov Decision Processes (MDPs), but also include a notion of 'terminal state' that does not affect expressivity (see subsection B.1).

Definition (Markov Decision Process)**.** Fix some set of states S, set of terminal states $S^+ \subseteq S$, initial state distribution $I \in \Delta(S)$, set of actions A and discount rate $\gamma \in (0, 1]$. Given any transition function $T : S \setminus S^+ \times A \to \Delta(S)$ and reward function $R : S \times A \times S \to \mathbb{R}$, we define an MDP $M_{T,R}$ as the tuple $(S, A, T, I, R, \gamma, S^+)$. Let $\mathcal{M}$ be the set of such MDPs. A memoryless policy $\pi : S \to \Delta(A)$, describes an agent's behaviour in any $M_{T,R}$ and Π denotes the set of these policies. A *trajectory*, $\tau \in S \times (A \times S)^*$, describes a "path" through any of the MDPs $M_{T,R} \in \mathcal{M}$ and consists of a sequence of states and actions. The return of a trajectory τ under reward function R is given by $G_R(\tau) = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})$. Given a policy π acting in $M_{T,R}$, the expected return from state s is given by: $V^{\pi}_{T,R}(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, T]$. We denote the set of optimal policies by $\text{opt}(M_{T,R}) \coloneqq \text{argmax}_{\pi \in \Pi} \mathbb{E}_{s_0 \sim I}[V^{\pi}(s_0) \mid \pi, T]$.

2 Distance-impact Penalties

Similarly to Krakovna et al. (2020b), we quantify the side effects of an agent's actions by comparing: (1) the way the world *is* after the agent has acted; and (2) the way the world *would have been*, had the agent never acted at all. Let $s_t \in S$ denote the state of the world at time t (after the agent has taken t actions) and $s^c_t \in S$ denote the *counterfactual state*: the way the world would have been if the agent had done nothing for t time steps. Finally, let $a_\times \in A$ denote a special "do-nothing" action. We can formally characterise s^c_t inductively: $s_0 \coloneqq s^c_0$ and $s^c_{t+1} \sim T(s_t, a_\times)$. Note that, when $T(s_t, a_\times)$ is stochastic, s^c_t may be a random variable. In this work, we assume it is deterministic and discuss this in appendices B.2.2 and B.2.3.

We quantify the impact of a sequence of t actions using the difference between s_t and s^c_t under a *state-distance measure*. Formally, a *state-distance measure* $d : S \times S \to \mathbb{R}_{\geq 0}$ is a function that quantifies the difference between two states where $d(x, x) = 0$ and $d(x, y) = d(y, x)$. The *impact* of an agent up to time t is given by $d(s_t, s^c_t)$. For example, one natural form for a state-distance measure would be $d(s, s') = f(|\phi(s) - \phi(s')|)$, where ϕ maps states to k-dimensional feature vectors, and f is a k-aray, monotonic function with codomain $\mathbb{R}_{\geq 0}$. Using our characterisation, we construct a reward signal that penalises actions that increase impact.

Definition (Δ_d)**.** Suppose d is a state-distance measure defined over state space S. Given action space A and a discount rate γ, define a distance-impact penalty $\Delta_d : S \times A \times S \to \mathbb{R}$:

$$\Delta_d(s_t, a_t, s_{t+1}) \coloneqq \gamma \cdot d(s_{t+1}, s^c_{t+1}) - d(s_t, s^c_t)$$

Given any MDP $M_{T,R}$, we can augment $M_{T,R}$ with the Δ_d to get $M_{T,R-\Delta_d}$, an MDP with reward:

$$(R - \Delta_d)(s_t, a_t, s_{t+1}) \coloneqq R(s_t, a_t, s_{t+1}) - \Delta_d(s_t, a_t, s_{t+1})$$

Distance-impact penalties *punish* actions that cause further impact when $d(s_{t+1}, s^c_{t+1}) > d(s_t, s^c_t)$ (up to time-discounting constant γ) and *reward* actions that decrease the amount of impact when $d(s_{t+1}, s^c_{t+1}) < d(s_t, s^c_t)$. Over the course of a trajectory, these rewards and punishments cancel out, giving a total punishment that depends only on the final state.

Theorem 1 (terminal distance)**.** *Fix any T, R and distance function d.*

(A) Consider any trajectory $\tau = s_0, a_0, \ldots, s_{|\tau|}$. Its return through $M_{T,R}$ and return through $M_{T,R-\Delta_d}$ satisfy $G_{R-\Delta_d}(\tau) = G_R(\tau) - \gamma^{|\tau|} \cdot d(s_{|\tau|}, s^c_{|\tau|})$

(B) Any policy $\pi : S \to \Delta(A)$ has initial state-values in $M_{T,R}$ and $M_{T,R-\Delta_d}$ that satisfy $V^{\pi}_{R-\Delta_d}(s_0) = V^{\pi}_R(s_0) - \mathbb{E}_\tau[\gamma^{|\tau|} \cdot d(s_{|\tau|}, s^c_{|\tau|})]$

Proof and a further result concerning infinite trajectories can be found in Appendix C. Intuitively, taking the summation of Δ_d terms "telescopes" the return, such that the only remaining terms are $d(s_0, s_0) = 0$ and $d(s_{|\tau|}, s^c_{|\tau|})$. The upshot of this theorem is that distance-impact penalties punish agents for exactly those side effects that persist once the agent has finished acting: the difference between the way the world ends up, and the way it would have ended up if the agent had not acted.

2.1 Analysing distance-impact penalties' effects on optimal behaviour

Ng et al. (1999) show that augmenting an MDP with potential-based reward functions of the form $F(s_t, a_t, s_{t+1}) = \gamma\Phi(s_{t+1}) - \Phi(s_t)$ does not change the set of optimal policies when the MDP has a single terminal state. Further, they show that this form of F is sufficient and necessary for this invariance when T and R are unknown. We restate this result in terms of distance measures and generalise it. In this section, we restrict ourselves to "static" environments (discussed further in subsection D.1) in which $\gamma = 1$, I is a degenerate distribution, and the inaction policy has no effect on the state (i.e. $T(s_0, a_\times) = s_0$ and thus $s^c_t = s_0$ for all t).

Proposition 1 (augmentation invariance)**.** When T and R are unknown, augmenting an MDP with a distance-impact penalty $-\Delta_d$ has no affect on optimal behaviour *if and only if* there exists some $c \in \mathbb{R}$ such that $d(s_+, s_0) = c$ for all $s_+ \in S^+$.

For a proof of this proposition, and a comparison with Ng et al. (1999) see subsection D.2. Intuitively, this implies that the addition of a distance-impact penalty usually creates some incentive toward lower-impact behaviour. One exception is in "goal-based" MDPs with a single, terminal "goal state": since $|S^+| = 1$, d is trivially constant over S^+ and so $-\Delta_d$ cannot affect behaviour.

The use of distance-impact penalties is motivated by the difficulty of designing a reward signal: when T is unknown, small misspecifications in R can lead to harmful behaviour. It is prudent, then, to consider how and whether misspecification in *distance measures* affects optimal behaviour. Suppose the measure d_1 accurately captures the (subjectively) important differences between states of the world, but our engineered measure d_2 differs from it slightly.

Theorem 2 (translation invariance)**.** *Say that two state-distance measures d_1 and d_2 are equivalent ($d_1 \equiv d_2$) if and only if augmenting any $M_{T,R} \in \mathcal{M}$ with $-\Delta_{d_1}$ induces the same set of optimal policies as augmenting it with $-\Delta_{d_2}$.*[2] *For d_1 and d_2 to be equivalent, it is sufficient and necessary that their difference is constant over terminal states:*

$$d_1 \equiv d_2 \iff \exists \delta \in \mathbb{R}, \forall s_+ \in S^+, d_1(s_+, s_0) - d_2(s_+, s_0) = \delta$$

For a proof, see subsection D.3. For a generalisation to known dynamics, see subsection D.4. One direction ($\Leftarrow$) of the theorem implies that misspecifying the state-distance measure by any constant amount will not change the optimal policies. The other direction ($\Rightarrow$) is more interesting: without prior knowledge of T or R, a constant difference is the only acceptable misspecification and any other change between two state-distance measures will impact learning. One important corollary is that distance-impact measures are not *scale-invariant*: in general, $d \not\equiv \mu \cdot d$ for $\mu \in \mathbb{R} \setminus \{1\}$. This allows us to weight R and Δ_d to trade off between completing the task and avoiding side effects. Fortunately, as the following result shows, distance-impact penalties are *less sensitive* to misspecification than reward functions.

Proposition 2 (ϵ-terminal difference)**.** If $\max_{s_+ \in S^+} |d_1(s_+, s_0) - d_2(s_+, s_0)| \leq \epsilon$, then for any π, we have $|V^\pi_{R-\Delta_{d_1}}(s_0) - V^\pi_{R-\Delta_{d_2}}(s_0)| \leq \epsilon$. Furthermore, an optimal policy π^*_2 for $R - \Delta_{d_2}$ is at most 2ϵ worse than optimal according to $R - \Delta_{d_1}$.

For a proof, see subsection D.5. Proposition 2 shows that, if we use an algorithm that converges to an optimal policy, then the cost of misspecifying d is at most 2ϵ. In contrast, misspecifying a reward function by at most ϵ on any s, a, s' can lead to a suboptimality of up to $2\epsilon|\tau|$ (see subsection D.6).

3 Empirical results

Using a series of grid-world environments and distance-impact measures, we empirically evaluate: (1) whether distance-impact penalties can be used to generate a variety of desirable behaviours; and (2)

[2] I.e., for all T, R, $\text{opt}(M_{T,R-\Delta_{d_1}}) = \text{opt}(M_{T,R-\Delta_{d_2}})$. We restrict consideration to transition functions that ensure termination (except for the $a_\times$ policy, which loops at s_0 forever).

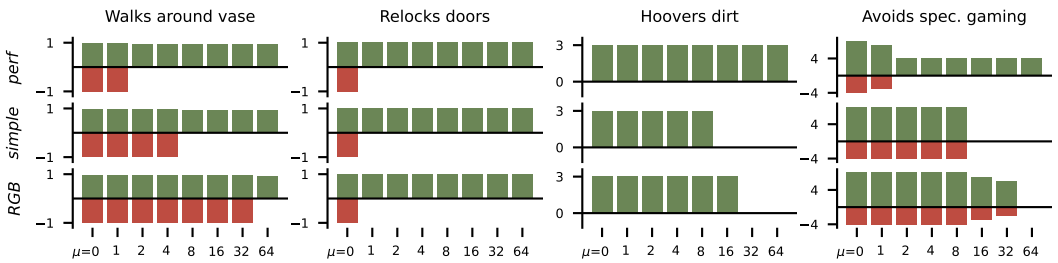

Figure 1: We trained RL agents until convergence on four grid-world environments, containing smashable vases, lockable doors, and dirt to be cleared. Green bars denote reward and red bars denote side effects. We evaluated three distance measures scaled by eight constants (μ), training agents with the reward $(R - \mu\Delta_d)$. The measure d_{perf} is a best-case distance measure, d_{simple} naively counts the differences between states, and d_{RGB} measures the Euclidean distance between RGB representations. All three are task-agnostic (shared between environments).

how robust the behaviour of the trained agent is to misspecification. More details about experiments and results can be found in Appendix E.

Appropriate state-distance measures create desirable behaviour. Our experiments demonstrate that a distance-impact penalty using an appropriate state-distance measure can prevent side effects whilst enabling task completion across a range of environments. Figure 1 shows that for $\mu \geq 2$, agents trained with distance measure $\mu \cdot d_{\text{perf}}$ learn to avoid unnecessary side effects (such as smashing vases), reverse their side effects (such as by re-locking doors), and avoid certain reward-hacking behaviours (such as smashing vases in order to clean them up), all while completing the task. In Figure 5 we show that, unlike naive regularisers, distance-impact measures can avoid "interference behaviour": acting in ways that prevent humans from having an impact on their environment (Krakovna et al., 2018). Figure 4 demonstrates that training a deep Q-network agent in a stochastic environment using distance-impact measures can promote lower-impact behaviour during training and in the final policy.

Performance is sensitive to distance specification. When μ is too low, agents will be overconfident and incur unnecessary side effects. When μ is too high, imperfect state-distance measures (such as d_{simple} and d_{RGB}) will induce overly cautious behaviour. This failure is graceful since a conservative engineer can overestimate μ, and an agent failing to complete a task but having no impact is preferable to catastrophic optimisation of the wrong objective. The performance of d_{RGB} demonstrates that there may be no value of μ that allows task completion and side effect avoidance across all tasks.

4 Discussion

Comparison to existing methods. Existing impact-regularisers, such as *Attainable Utility Preservation (AUP)* (Turner et al., 2020) and *Future Tasks (FT)* (Krakovna et al., 2020b) focus on option-value: they are naturally biased towards discouraging the side effects an agent could not later fix. In contrast, distance-impact measures don't inherently discriminate between reversible and irreversible side effects (except when the agent *really does* reverse the effect) and may therefore be oversensitive to mundane, reversible changes (such as moving a chair). However, FT and AUP may ignore side effects that are reversible but dangerous: for example, unlocking a door is a reversible action, but may leave people or property insecure. Further, unlike AUP and FT, the distance-impact is *not* sensitive to state dynamics, allowing for generalisation between environments and evaluation when they are *unknown*. All three regularisers are scope-sensitive and avoid interference. Unlike AUP, distance-impact penalties and FT penalise delayed side effects but do *not* penalise power-seeking behaviour (which is arguably beyond the scope of side effect avoidance). For a table comparing regularisers' properties, see Appendix G.

Future work. More work could be done to formalise the notion of a counterfactual world-state when T is stochastic with respect to $a_\times$ (see B.2.3). When state dynamics are unknown, FT and AUP's penalties must be learnt via Q-functions; further work could also compare the *sample efficiency* and *computational complexity* of these methods. There are also more principled ways to generate distance measures (such as via human feedback).

References

Parand Alizadeh Alamdari, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. 2021. Avoiding Negative Side Effects by Considering Others. In *Safe and Robust Control of Uncertain Systems Workshop at NeurIPS 2021*. Safe and Robust Control of Uncertain Systems Workshop at NeurIPS 2021. https://sites.google.com/view/safe-robust-control/home

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. http://arxiv.org/abs/1606.06565 arXiv: 1606.06565.

Stuart Armstrong and Benjamin Levinstein. 2017. Low impact artificial intelligences. https://doi.org/10.48550/arXiv.1705.10720

Carla Zoe Cremer and Luke Kemp. 2021. *Democratising Risk: In Search of a Methodology to Study Existential Risk*. Technical Report. University of Oxford - Future of Humanity Institute; University of Oxford - Medical Sciences Division & Centre for the Study of Existential Risk, University of Cambridge.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2022. Unsolved Problems in ML Safety. http://arxiv.org/abs/2109.13916 arXiv:2109.13916 [cs].

Dan Hendrycks and Mantas Mazeika. 2022. X-Risk Analysis for AI Research. https://doi.org/10.48550/arXiv.2206.05862 arXiv:2206.05862 [cs].

Victoria Krakovna, Uesato Jonathan, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. 2020a. Specification gaming: the flip side of AI ingenuity. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. 2018. Penalizing side effects using stepwise relative reachability. https://doi.org/10.48550/ARXIV.1806.01186

Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. 2020b. Avoiding Side Effects By Considering Future Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 19064–19074. https://proceedings.neurips.cc/paper/2020/file/dc1913d422398c25c5f0b81cab94cc87-Paper.pdf

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (Feb. 2015), 529–533. https://doi.org/10.1038/nature14236 Number: 7540 Publisher: Nature Publishing Group.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. 1999. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27 - 30, 1999*, Ivan Bratko and Saso Dzeroski (Eds.). Morgan Kaufmann, 278–287.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (Jan. 2016), 484–489. https://doi.org/10.1038/nature16961 Number: 7587 Publisher: Nature Publishing Group.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and Characterizing Reward Hacking. http://arxiv.org/abs/2209.13085 arXiv:2209.13085 [cs, stat].

Alex Turner, Neale Ratzlaff, and Prasad Tadepalli. 2020. Avoiding Side Effects in Complex Environments. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 21406–21415. https://proceedings.neurips.cc/paper/2020/file/f50a6c02a3fc5a3a5d4d9391f05f3efc-Paper.pdf

Peter Vamplew, Cameron Foale, Richard Dazeley, and Adam Bignold. 2021. Potential-based multiobjective reinforcement learning approaches to low-impact agents for AI safety. *Engineering Applications of Artificial Intelligence* 100 (April 2021), 104186. https://doi.org/10.1016/J.ENGAPPAI.2021.104186 Publisher: Pergamon.

A Analysis of relation to existential risk reduction

Following (Hendrycks and Mazeika, 2022), we characterise existential risks as those that can permanently curtail humanity's long-term potential. Although "potential" is a relative and value-laden term (see Cremer and Kemp (2021)), for this analysis we will assume that the extinction or domination of humanity by advanced AI systems would be extremely bad. We believe distance-impact measures contribute to existential risk reduction through alignment-research: that "which seeks to make AI systems less hazardous by focussing on hazards such as power-seeking tendencies, dishonesty or hazardous goals" (Hendrycks and Mazeika, 2022). In particular, we argue that distance-impact measures contribute to the reduction of harm from *proxy misspecification*: the optimisation of faulty proxy objectives.

To see how distance-impact measures might reduce harm, suppose an engineer designs a proxy reward to approximate some "ideal" reward. (Whether there is some "ideal" reward function that precisely captures the engineer's values is a contentious moral issue and whether it is possible to capture "humanity's" values is even more so. For now, we suppose such an ideal reward exists.) In complex environments, the perfect reward and engineered reward function will differ (Skalse et al., 2022). Crucially, competent optimisation of the engineered reward may be significantly worse than doing nothing at all. This superhuman agent could have significant unintended side effects, including effects that curtail humanity's potential (e.g. through extinction). Even though adding an impact-regulariser may not help us capture the "ideal" reward, it may give a proxy that is less prone to creating catastrophic behaviour. It will fail in ways that have lower impact, and therefore fewer catastrophic events.

One possible benefit is that an engineer that uses an impact regulariser may be able to improve their control over the agent's behaviour. Rather than specifying a complex reward that attempts to capture all human values, the engineer can design a simple "instruction-like" reward and then augment it with the impact regulariser. For example, a simple function that gives 1 reward exactly when a robot hands its owner milk bought from the grocery store can be augmented with an impact regulariser to get an agent that might be trusted to act autonomously without incurring the side effects of walking into people or messing up the grocery store. In contrast, suppose an agent is rewarded for giving its owner milk, and penalised whenever an agent is hurt or the grocery store becomes messy. This agent may take extreme action to avoid these rewards (perhaps by taking control of the grocery store to avoid any mess).

There are a couple of immediate limitations to distance-impact penalties for reducing catastrophic risk. First, distance-impact penalties may be less suitable than other impact regularisers when it comes to tail risks since they don't focus on irreversible actions. In particular, most existential risk is by definition irreversible. However, distance-impact penalties may be much more scalable, and therefore more applicable in practice. Second, like other impact-regularisers, distance-impact penalties may be difficult to apply in complex environments with partial information (such as the real world). In particular, creating a tamper-proof system that can evaluate the state and counterfactual state and feed those to an advanced AI system may be difficult.

B Details on formalism

B.1 MDPs with terminal states

Recall that our MDPs take the form $(S, A, T, I, R, \gamma, S^+)$, where S^+ is a set of terminal states. Although it is not ubiquitous to define an MDP using a set of terminal states S^+, it is not restrictive to do so. Defining a non-empty set S^+ is equivalent to having a standard set of states S and infinite trajectories: for any state $s_+ \in S^+$, set $T(s_+ \mid a, s_+) = 1$ for all a, and fix the reward to 0. Note that we would have to redefine Δ_d such that for any $s_+ \in S^+$, $\Delta_d(s, a, s') = 0$. Similarly, any infinite-horizon MDP can be expressed as an MDP with an empty set of terminal states $S^+ = \emptyset$. Further, having a set of S^+ can be quite natural in some cases. Suppose we have some MDP $(S, A, T, I, R, \gamma, \_)$ and we want to fix episodes to have a length of 100. We include in our state definition some timer ($S' = S \times \{0, \ldots, 100\}$) and then choose $S^+ = S \times \{100\}$.

B.2 Counterfactuals and ambiguity

There are three complications in the definition of the counterfactual state s_t^c concerning how to define a "do nothing" action, how to ensure $-\Delta_d$ is non-Markovian, and how to ensure s_t^c is well-defined in counterfactual environments. To the best of our knowledge, all previous work on side-effects that uses counterfactuals either shares these complications (Krakovna et al., 2020b) or avoids them by only considering cases where $s_t^c = s_0$ (Alamdari et al., 2021). In this section, we describe each problem and propose solutions.

B.2.1 A "do nothing" action

Our characterisation of impact requires a privileged "do-nothing" action $a_\times$. However, for some states of the world, there is no appropriate "do-nothing" action. Suppose an agent is driving a car with several passengers (Krakovna et al., 2020b). The closest thing to not acting could be to take one's foot off the pedals and release the steering wheel, but such an action would be dangerous. To some extent, this problem is avoided for initial inaction counterfactuals so long as there is some $a_\times$ action in at least the initial state: if the initial state is one in which the car is parked, then $a_\times$ is well defined. We can therefore maintain that s_t^c is well defined for the initial inaction counterfactual.

B.2.2 Non-Markovian rewards

For many environments, the counterfactual state is under-determined by the current state and therefore the distance-impact penalty might be non-Markovian. A Markovian reward is a function of the state s_t, action a_t and next state s_{t+1}. Consider again our augmented reward function:

$$R'(s_t, a_t, s_{t+1}) = R(s_t, a_t, s_{t+1}) - \big(\gamma \cdot d(s_{t+1}, s_{t+1}^c) - d(s_t, s_t^c)\big)$$

The terms in our reward function concern the counterfactual states s_t^c and s_{t+1}^c which may not be a function of (s_t, a_t, s_{t+1}). Two possible reasons for this are:

Stochastic initial state. The distance-impact penalty can be non-Markovian when the initial state is stochastic. Suppose an agent operates in an environment containing some lights, that are sometimes on and sometimes off at initialisation. Given only information about the current time step in which the lights are on, it is unclear whether the agent will receive a reward or penalty for turning them off: whether the lights were off or on in the original state.

Unknown time step. The distance-impact penalty can also be non-Markovian when the initial state is fixed (or known) but the time step is not known. If some part of the environment changes over time regardless of whether the agent acts, then knowing what the current time step is will be essential for knowing how it would have been had the agent never acted.

In deterministic environments, there are two ways to recover the Markov property. First, by restricting to "static" environments with deterministic start-states (i.e. the $a_\times$ action changes nothing, and $I(s_0) = 1$). In these environments, $s_t^c = s_0$ is constant across all trajectories and times and therefore the penalty $\Delta_d(s_t, a_t, s_{t+1}) = \gamma \cdot d(s_{t+1}, s_0) - d(s_t, s_0)$ is Markovian. This is the strategy we follow in subsection 2.1.

Another strategy, which applies to any MDP where $T(a_\times, s)$ deterministic, would be to create an augmented MDP with state-space $S' = S \times \mathbb{N} \times S$, where $s'_t = (s_t, t, s_0)$. When the starting state

(s_0) and time step (t) are known, s^c_{t+1} is entirely determined by the deterministic series:

$$\begin{aligned} s^c_0 &= s_0 \\ s^c_t &= T(s^c_1, a_\times) \\ &\vdots \\ s^c_{t+1} &= T(s_t, a_\times) \end{aligned}$$

Since the counterfactual state is now wholly determined for any state $s' = (s, t, s_0)$, the distance impact penalty Δ_d is a Markovian reward.

B.2.3 Stochastic dynamics

In environments with a stochastic transition function, including the initial state and time step is not sufficient for determining a single counterfactual state. If $T(s, a_\times)$ is a probability distribution with support over multiple states, the counterfactual state s^c_t becomes poorly defined. For example, suppose that there is a cat in a box that has a 50:50 chance of being poisoned if the agent does not act. If the agent takes the cat out of the box, should we evaluate its impact with respect to the world in which there is an alive cat in a box, or the world in which there is a dead cat in a box? Formally, if the inaction policy ($\pi_{no-op}(s) = a_\times$) acting in an MDP M induces a distribution over trajectories in which $\mathbb{P}_{\pi_{no-op}}[\tau] < 1$ for all trajectories τ, then at any given time, their are multiple possible counterfactual states.

In this work, we avoid this problem by restricting to MDPs in which T is deterministic with respect to the $a_\times$ action but can be stochastic otherwise.

C Further proofs and results

C.1 Proof of Theorem 1

Proof of (A).
For any reward R and penalty Δ_d, consider a trajectory τ defined for M_R and $M_{R-\Delta_d}$. We can separate the trajectory's return into two components:

$$\begin{aligned} G_{R-\Delta_d}(\tau) &= \sum_{t=0}^{|\tau|-1} \gamma^t \cdot (R(s_t, a_t, s_{t+1}) - \Delta_d(s_t, a_t, s_{t+1})) && \textit{(definition of return)} \\ &= G_R(\tau) + G_{-\Delta_d}(\tau) && \textit{(as above)} \end{aligned}$$

We then simplify the second term:

$$\begin{aligned} G_{-\Delta_d}(\tau) &= -\sum_{t=0}^{|\tau|-1} \gamma^t \cdot (\gamma \cdot d(s_{t+1}, s^c_{t+1}) - d(s_t, s^c_t)) && \textit{(definition of } \Delta_d\textit{)} \\ &= \sum_{t=0}^{|\tau|-1} \gamma^t \cdot d(s_t, s^c_t) - \gamma^{t+1} \cdot d(s_{t+1}, s^c_{t+1}) && \textit{(algebra)} \end{aligned}$$

One can expand the summation in this equation:

$$(\gamma^0 d(s_0, s^c_0) - \gamma^1 d(s_1, s^c_1)) + (\gamma^1 d(s_1, s^c_1) - \gamma^2 d(s_2, s^c_2)) + \dots \\ \dots + (\gamma^{|\tau|-1} d(s_{|\tau|-1}, s^c_{|\tau|-1}) - \gamma^{|\tau|} d(s_{|\tau|}, s^c_{|\tau|})) \quad (1)$$

Adjacent terms, share the same exponent for γ and cancel out:

$$\gamma^0 d(s_0, s^c_0) + \cancel{-\gamma^1 d(s_1, s^c_1) + \gamma^1 d(s_1, s^c_1)} + \cancel{-\gamma^2 d(s_2, s^c_2) +} \dots \\ \dots \cancel{+\gamma^{|\tau|-1} d(s_{|\tau|-1}, s^c_{|\tau|-1})} - \gamma^{|\tau|} d(s_{|\tau|}, s^c_{|\tau|}) \quad (2)$$

We can additionally cancel the first term, since $s_0^c = s_0$ and $d(s, s) = 0$. This leaves us with the final expression:

$$G_{R-\Delta_d}(\tau) = G_R(\tau) - \gamma^{|\tau|} \cdot d(s_{|\tau|}, s^c_{|\tau|})$$

□

Proof of (B).
For any reward R and penalty Δ_d, consider a policy π defined for M_R and M_{Δ_d} and the state-value function for an initial state:

$$\begin{aligned}
V^\pi_{R-\Delta_d}(s_0) &= \mathop{\mathbb{E}}_{\tau}[\sum_{t=0}^{|\tau|} \gamma^t (R - \Delta_d)(s_t, a_t, s_{t+1})] && \textit{(definition of } V\textit{)} \\
&= \mathop{\mathbb{E}}_{\tau}[G_{R-\Delta_d}(\tau)] && \textit{(definition of } G\textit{)} \\
&= \mathop{\mathbb{E}}_{\tau}[G_R(\tau) - \gamma^{|\tau|} \cdot d(s_{|\tau|}, s^c_{|\tau|})] && \textit{(Part (A))} \\
&= V^\pi_R(s_0) - \mathop{\mathbb{E}}_{\tau}[\gamma^{|\tau|} \cdot d(s_{|\tau|}, s^c_{|\tau|})] && \textit{(L.O.E. and def. of } V_R\textit{)}
\end{aligned}$$

□

C.2 Additional result: without termination, impact doesn't matter

Distance-impact measures, as defined in section 2, are only effective for finite trajectories. Intuitively, we penalise the agent for its final impact, and there is no such impact when an agent never halts.

Proposition 3 (Non-terminal distance)**.** Fix any S, A, T, I, γ, and S^+. For any R and distance penalty Δ_d, if the MDPs M_R and $M_{R-\Delta_d}$ are non-terminating then:

(A) An infinite trajectory τ through $M_{R-\Delta_d}$, has return given by:
$$G_{R-\Delta_d}(\tau) = G_R(\tau)$$

(B) A policy π for $M_{R-\Delta_d}$, has initial-state value given by:
$$V^\pi_{R-\Delta_d}(s_0) = V^\pi_R(s_0)$$

Proof of Proposition 3. For any infinite trajectory τ, consider return of the τ through $M_{R-\Delta_d}$. Proof that $G_{R-\Delta_d}(\tau) = G_R(\tau) + G_{-\Delta_d}(\tau)$ is identical to that given for Theorem 1.(A). Therefore it suffices to shows that $G_{\Delta_d}(\tau) = 0$:

$$\begin{aligned}
G_{\Delta_d}(\tau) &= \sum_{t=0}^{\infty} \gamma^t \cdot (-\Delta_d)(s_t, a_t, s_{t+1}) && \textit{(definition of return)} \\
&= -\sum_{t=0}^{\infty} \gamma^t \cdot (\gamma \cdot d(s_{t+1}, s^c_{t+1}) - d(s_t, s^c_t)) && \textit{(algebra)} \\
&= \sum_{t=0}^{\infty} \gamma^t \cdot d(s_t, s^c_t) - \sum_{t'=1}^{\infty} \gamma^{t'} \cdot d(s_{t'}, s^c_{t'}) && \textit{(algebra and } t' = t+1\textit{)} \\
&= \gamma^0 \cdot d(s_0, s^c_0) && \textit{(cancel terms)} \\
&= 0 && (d(s_0, s^c_0) = 0)
\end{aligned}$$

Proof of part B is nearly identical to proof of Theorem 1.(B).

□

In practice, almost all environments are episodic, as it is impossible to train an agent for infinitely long. Still it remains an interesting question for future work to see if low-impact can be formalised in infinite trajectories.

D Sensitivity and Invariance

This appendix contains detailed and expanded explanation of the results in subsection 2.1 as well as proofs of the results.

D.1 Restricting to 'static' environments

In subsection 2.1, we assume MDPs are episodic following the concerns about infinite trajectories outlined in subsection B.2. We also assumed that environments were 'static', meaning $a_\times$ is well-defined, that the environment is static with respect to $a_\times$ ($T(s \mid s, a_\times) = 1$) and that there is some assume there fixed initial state (s_0). These assumptions give us the desirable property that $s_t^c = s_0$ for all time steps t. This is sufficient (but not necessary) for the penalty to be Markovian, and also allows us to reduce the distance measure to consider d in terms of the unary function $d(\cdot, s_0)$.

This class of MDPs captures a fairly large range of natural environments: any that involves a single agent operating in an environment without significant automatic processes (such as conveyor belts). Further, considering static environments will allow me to draw a tight correspondence with Ng et al. (1999). Note that in the unary case, distance-impact penalties are similar to Vamplew et al. (2021), which performs impact regularisation in the lexicographic objective setting.

D.2 Invariances to the addition of a distance measure

Ng et al. (1999) shows that the addition of certain kinds of reward functions to an original reward does not change which policies are optimal. In particular potential-based reward functions of the form $F(s_t, a_t, s_{t+1}) = \gamma\Phi(s_{t+1}) - \Phi(s_t)$.

Theorem 3 (Ng et al. (1999))**.** *Fix any S, A, s_0, $\gamma = 1$, and S^+ with a single terminal state ($S^+ = \{s_+\}$). Given any auxiliary reward F:*

$$F \textit{ is a potential-based reward function}$$

$$\iff$$

$$\text{opt}(M_{T,R}) = \text{opt}(M_{T,R+F}) \textit{ for all } T, R$$

Adding a potential-based shaping function does not change the optimal policy but any other function will. Crucially, since $d(\cdot, s_0)$ is a potential function and $-\Delta_d = \gamma \cdot d(s_{t+1}, s_0) - d(s_t, s_0)$ is a potential shaping function, adding a distance-impact penalty does not change the optimal behaviour *if there is a single terminal state*.

Corollary 3.1 (single-state invariance)**.** Fix any S, A, s_0, $\gamma = 1$, and S^+ with a single terminal state ($S^+ = \{s_+\}$). Given any state-distance measure d:

$$\text{opt}(M_{T,R}) = \text{opt}(M_{T,R-\Delta_d}) \text{ for all } T, R$$

When there is a single terminal state, a distance-impact penalty has no effect on behaviour. This is to be expected since in such cases there is a fixed outcome and therefore the agent cannot *impact* the outcome. Regularisers are concerned with environments in which an agent can have *more or less* impact and therefore there are multiple possible outcomes. Proposition 1 generalises this result to the multiple-terminal state case to consider when adding a distance-impact penalty affects behaviour.

Like Ng et al. (1999), we have assumed $\gamma = 1$. When $\gamma < 1$, all non-zero distance measures affect optimality for some transition function. (*To see this, set $R = 0$ and use the transition function T that self-loops on action a_1 with $\mathbb{P} = 0.99$ and goes to some s_+ with $d(s_+, s_0) > 0$ otherwise and for any other action. Note that* $\text{opt}(M_{T,0}) = \Pi$*, but that* $\text{opt}(M_{T,0-\Delta_d}) = \{\pi\}$*, where π is the policy that always chooses a_1: maximising the trajectory length and minimising $\gamma^{|\tau|} \cdot D(s_0)$.*)

Proof of Proposition 1 is a special case Theorem 2: setting $d_1 = d$ and $d_2(s, a, s') = 0$ tells us that adding $-\Delta_d$ is impactful only when $(d_1 - d_2) = d_1$ is constant over the terminal states. These are sufficient and necessary conditions for optimal behaviour to be invariant to the inclusion of a distance-impact penalty (when $\gamma = 1$). Theorem 3 is a special case of Proposition 1 as the distance-impact penalty must be constant over a singleton set of terminal states.

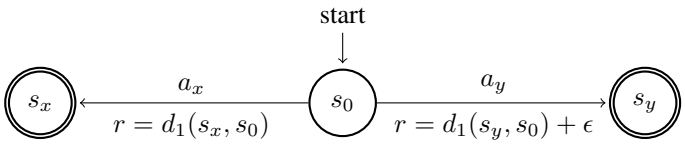

Figure 2: A depiction of the MDP $M_{T,R}$.

D.3 Proof of Theorem 2

Proof of Theorem 2 ($\Leftarrow$). Suppose $(d_1 - d_2)(\cdot, s_0)$ has constant value δ over terminal states S^+. Then for any policy π:

$$\begin{aligned} V^\pi_{R-\Delta_{d_1}}(s_0) &= V^\pi_R(s_0) - \mathop{\mathbb{E}}_{\tau}[\gamma^{|\tau|} \cdot d_1(s_{|\tau|}, s_0)] && \textit{(Theorem 1)} \\ &= V^\pi_R(s_0) - \mathop{\mathbb{E}}_{\tau}[\gamma^{|\tau|} \cdot d_2(s_{|\tau|}, s_0))] - \delta && \textit{(by assumption)} \\ &= V^\pi_{R-\Delta_{d_1}}(s_0) - \delta \end{aligned}$$

Therefore $V^{\pi_1}_{R-\Delta_{d_1}}(s_0) \geq V^{\pi_2}_{R-\Delta_{d_1}}(s_0)$ iff $V^{\pi_1}_{R-\Delta_{d_2}}(s_0) \geq V^{\pi_2}_{R-\Delta_{d_2}}(s_0)$ and d_1 and d_2 give the same optimal policies. □

Constructive proof of Theorem 2 ($\Rightarrow$). Fix any S, A, $\gamma = 1$, s_0, S^+ and two distance measures d_1 and d_2 such that $(d_1 - d_2)(\cdot, s_0)$ differs over S^+. Let s_x and s_y be any two terminal states over which $(d_1 - d_2)(\cdot, s_0)$ differs.

We choose T and R to construct a determinstc MDP $M_{T,R}$ (depicted in Figure 2) in which only three states are reachable: s_0, s_x and s_y. Let a_x be any action in A and define the transition function T as follows:

$$T(s_0, a) = \begin{cases} s_x & \text{if } a = a_x \\ s_y & \text{otherwise} \end{cases}$$

Since T is deterministic, an agent has two choices: move to s_x, or move to s_y. We now construct the reward function R:

$$R(s_0, a, s') = \begin{cases} d_1(s_x, s_0) & \text{if } s' = s_x \\ d_1(s_y, s_0) & \text{if } s' = s_y \end{cases}$$

To show that $d_1 \not\equiv d_2$, it will be sufficient to show that $\text{opt}(M_{T,R-\Delta_{d_1}}) \neq \text{opt}(M_{T,R-\Delta_{d_2}})$. We do this by first showing that $\text{opt}(M_{T,R-\Delta_{d_1}}) = \Pi$, and then showing that some policy is not in $\text{opt}(M_{T,R-\Delta_{d_2}})$.

Let a_y be any action in $A \setminus \{a_x\}$ and note that there are only two possible trajectories in either MDP (up to action-equivalence):

$$\tau_x = s_0, a_x, s_x$$
$$\tau_y = s_0, a_y, s_y$$

By Theorem 1 we can see both, and therefore all, trajectories get 0 return:

$$G_{R-\Delta_{d_1}}(\tau_x) = R(s_0, a_x, s_x) - d_1(s_x, s_0) = 0$$
$$G_{R-\Delta_{d_1}}(\tau_y) = R(s_0, a_y, s_y) - d_1(s_y, s_0) = 0$$

Therefore, for any policy π, $V^\pi_{R-\Delta_{d_1}}(s_0) = 0$. It follows that all policies are optimal: $\text{opt}(M_{T,R-\Delta_{d_1}}) = \Pi$.

It remains to show our second, claim that there is a suboptimal policy for d_2. Let π_x and π_y be the policies that always choose a_x and a_y respectively. Consider their value functions in the second MDP, $M_{T,R-\Delta_{d_2}}$:

$$V^{\pi_x}_{R-\Delta_{d_2}}(s_0) \quad = d_1(s_x, s_0) - (d_2(s_x, s_0) - d_2(s_0, s_0)) = (d_1 - d_2)(s_x, s_0)$$
$$V^{\pi_y}_{R-\Delta_{d_2}}(s_0, s_0) = d_1(s_y, s_0) - (d_2(s_y, s_0) - d_2(s_0, s_0)) = (d_1 - d_2)(s_y, s_0)$$

By assumption $(d_1 - d_2)(s_x, s_0) \neq (s_1 - s_2)(s_y, s_0)$, and therefore π_x and π_y have different state-value functions. It follows that one of these policies is sub-optimal, that $\text{opt}(M_{T,R-\Delta_{d_2}}) \neq \Pi$ and therefore that $d_1 \not\equiv d_2$.

□

D.4 Invariance under known dynamics

The first direction ($\Rightarrow$) of Theorem 2 implies that misspecifying the distance measure by any constant amount will not change the optimal policies. The second direction ($\Leftarrow$) is more interesting: that without further assumptions about the nature of either T or R, this is the only acceptable misspecification. Any other change between two distance measures will impact learning.

The necessity claim is equivalent to saying that, for at least one T, there is some R for which d_1 and d_2 will disagree. However, a stronger claim is also true: for almost *any* T, there is some R for which d_1 and d_2 disagree.

Definition ($\equiv_T$)**.** For any fixed S, A, $\gamma = 1$, s_0, S^+, and transition function T, say that two distance measures d_1 and d_2 are T-equivalent ($d_1 \equiv_T d_2$) if and only if for any R, $\text{opt}(M_{T,R-\Delta_{d_1}}) = \text{opt}(M_{T,R-\Delta_{d_2}})$.

Remark. Note that $d_1 \equiv d_2$ if and only if $\forall T, d_1 \equiv_T d_2$.

To make an interesting claim about $\equiv_T$, we have to rule out a few edge cases. For example, MDPs with transition functions that are insensitive to rewards and deprive the agent of true "agency". The following condition rules out cases where no agent has control over its impact:

Definition (The agency condition)**.** Say that a rewardless MDP $M_{T,0}$ and distance measure d satisfy the *agency condition* when agents have at least some control over their terminal impact according to d. That is, there are two policies π_1 and π_2 that have different expected terminal impact:

$$\mathop{\mathbb{E}}_{\tau \sim Traj(\pi_1, M)}[d(s_\tau, s_0)] \neq \mathop{\mathbb{E}}_{\tau \sim Traj(\pi_2, M)}[d(s_\tau, s_0)]$$

The *agency condition* effectively says that policies, in principle, have some agency over how much impact they have: a fairly minor condition.

Theorem 4 (Translation Agency)**.** *Suppose $\gamma = 1$, T is any transition function, and d_1 and d_2 are distance measures:*

$$d_1 \equiv_T d_2 \iff M_{T,0} \textit{ fails to satisfy the agency condition for } (d_1 - d_2)$$

Before we prove this theorem, it will pay to look at a more intuitive, but less general result, restricted to deterministic policies.

Corollary 4.1 (Deterministic Translation Invariance)**.** Suppose $\gamma = 1$, T is a deterministic transition function, and let $S^+_T \subseteq S^+$ be the terminal states that are reachable with respect to T.

$$d_1 \equiv_T d_2 \iff d_1 - d_2 \text{ is constant over } S^+_T$$

If T is deterministic, and $(d_1 - d_2)$ is not constant over the *reachable* terminal states in S^+. The agent can choose which of the reachable terminal states it ends up in, and can therefore vary its impact w.r.t $(d_1 - d_2)$ and T satisfies the agency conditions.

One direction ($\Leftarrow$) of the proof of Theorem 4, is extremely similar to subsection D.3. The second is a proof that for most MDPs (all that satisfy the *agency condition*), optimality of behaviour is invariant only to translation (over the terminal states).

Theorem 4 ($\Rightarrow$). This proof is similar to that of Theorem 2, except that we cannot force trajectories to have length 1. Given any transition function T, and two measures d_1 and d_2, construct a reward function R that will separate optimal behaviour for d_1 and d_2:

$$R(s,a,s') = \begin{cases} 0 & \text{if } s' \notin S^+ \\ d_1(s', s_0) & \text{if } s' \in S^+ \end{cases}$$

Consider any trajectory τ through $M_{T,R-\Delta_{d_1}}$. By Theorem 1, the reward received from entering a terminal state exactly counters the distance-impact penalties up to reaching that terminal state.

$$\begin{aligned} G_{R-\Delta_{d_1}}(\tau) &= 0 + \ldots + 0 + R(s_{|\tau|-1}, a_{|\tau|-1}, s_{|\tau|}) - d_1(s_{|\tau|}, s_0) \\ &= d_1(s_{|\tau|}, s_0) - d_1(s_{|\tau|}, s_0) && = 0 \end{aligned}$$

It follows from this that *all* policies are optimal for $M_{T,R-\Delta_{d_1}}$:

$$\text{opt}(M_{T,R-\Delta_{d_1}}) = \Pi$$

Consider, in contrast, the return of any trajectory using d_2:

$$\begin{aligned} G_{R-\Delta_{d_2}}(\tau) &= 0 + \ldots + 0 + R(s_{|\tau|-1}, a_{|\tau|-1}, s_{|\tau|}) - d_2(s_{|\tau|}, s_0) \\ &= d_1(s_{|\tau|}, s_0) - d_2(s_{|\tau|}, s_0) && = d_1 - d_2)(s_{|\tau|}, s_0) \end{aligned}$$

The state-value function for a policy π in $M_{T,R-\Delta_{d_2}}$ is given by:

$$V^{\pi}_{R-\Delta_{d_2}}(s_0) = \mathop{\mathbb{E}}_{\tau \sim Traj(\pi,M)}[(d_1 - d_2)(s_\tau)]$$

Since every policy is optimal for $M_{T,R-\Delta_{d_1}}$, $\text{opt}(M_{T,R-\Delta_{d_2}}) = \text{opt}(M_{T,R-\Delta_{d_1}})$ is optimal if and only if every policy is optimal in $M_{T,R-\Delta_{d_1}}$. This is true exactly when the value $\mathbb{E}_{\tau \sim Traj(\pi,M)}[(d_1 - d_2)(s_\tau)]$ is constant between policies and T fails to satisfy the agency conditions w.r.t $(d_1 - d_2)$. Therefore, if T satisfies the agency conditions, then $\text{opt}(M_{T,R-\Delta_{d_2}}) \neq \text{opt}(M_{T,R-\Delta_{d_1}})$ and therefore $d_1 \not\equiv d_2$ □

Theorem 4 says that (in most natural environments), any variation in distance measures will lead to variation in optimality for at least some reward function. For example, suppose the engineer programs an agent to work in a museum but leaves its reward variable so that she can assign it to do a number of tasks: clean, patrol or move objects. Suppose further that she designed two candidate distance-impact measures that differ over more than translation. Theorem 4 demonstrates that there will be some tasks she could ask the agent to complete for which its optimal behaviour will differ when she switches distance measures.

D.5 Proof of Proposition 2

Proof. If $\gamma = 1$ or episode length is fixed to $|\tau|$, then the terminal discount factor $\gamma^{|\tau|}$ is constant: call it c. By Theorem 1, $V^{\pi}_{R-\Delta_D}(s_0) = V^{\pi}_{R}(s_0) - \mathbb{E}_\tau[c \cdot D(s_{|\tau|})]$. When calculating the difference between $V^{\pi}_{R-\Delta_{d_1}}(s_0)$ and $V^{\pi}_{R-\Delta_{d_2}}(s_0)$, the $V^{\pi}_{R}(s_0)$ terms will cancel out giving:

$$\begin{aligned} \left|V^{\pi}_{R-\Delta_{d_1}}(s_0) - V^{\pi}_{R-\Delta_{d_2}}(s_0)\right| &= \left|c \cdot \mathbb{E}\left[d_1(s_{|\tau|}, s_0) - c \cdot d_2(s_{|\tau|}, s_0)\right]\right| && \textit{(By Theorem 1)} \\ &\leq c \cdot \mathbb{E}\left[\left|d_1(s_{|\tau|}, s_0) - d_2(s_{|\tau|}, s_0)\right|\right] && \textit{(Jensen's inequality)} \\ &\leq \epsilon \end{aligned}$$

The final line follows from the assumption that $|d_1 - d_2| \leq \epsilon$, and therefore the inside of the expectation and the expectation itself are bounded by ϵ. Since $c \leq 1$, the difference in state-value is bounded ϵ.

Suppose π_2 is an optimal policy for $R - \Delta_{d_2}$ and π_1 is an optimal policy for $R - \Delta_{d_1}$. Then consider π_2's performance with respect to $R - \Delta_{d_1}$:

$$
\begin{aligned}
V^{\pi_2}_{R-\Delta_{d_1}}(s_0) &\geq V^{\pi_2}_{R-\Delta_{d_2}}(s_0) - \epsilon && \textit{(By the above)} \\
&\geq V^{\pi_1}_{R-\Delta_{d_2}}(s_0) - \epsilon && \textit{(Since } \pi_2 \textit{ is optimal for } \Delta_{d_2}) \\
&\geq V^{\pi_1}_{R-\Delta_{d_1}}(s_0) - 2\epsilon && \textit{(By the above)}
\end{aligned}
$$

Therefore, π_2 is at most 2ϵ worse than any other policy under $R - \Delta_{d_1}$. □

D.6 Consequences of misspecifying a reward function.

Shortly after Proposition 2, we mentioned that misspecifying a reward function by at most ϵ can lead to an optimality of at least $|\tau| \cdot 2\epsilon$. To see this, for any $n \in N$ construct two MDPs $M^n_{R_1}$ and $M^n_{R_2}$ with states $\{0, \ldots, n\}$, actions a_1 and a_2, one terminal state (n), one initial state (0) and $\gamma = 1$. Let $T(i, a) = i + 1$. Therefore, all trajectories through the MDPs have length n. Finally, let $R_1(s, a, s') = 0$ and $R_2(s, a_1, s') = 1$ but $R_2(s, a_2, s') = 0$. All policies get expected return 0 in $M^n_{R_1}$ and are therefore optimal. This includes the policy $\pi_2(s) = a_2$. However, π_2 gets return 0 in $M^n_{R_2}$, which is $2n\epsilon$ less than the policy $\pi_1(s) = a_1$.

E Experiments

We ran a series of experiments using grid world environments. A multi-purpose robot called Rob works in a museum containing: walls (square), goals (star), some dirt (circle), a vase (vase) or a locked door that can be unlocked (padlock). See Figure 3 for a depiction of the environments. The agents were fed RGB representations of the gridworld, with one pixel for each cell. These images are distinct from the depictions shown in this paper.

E.1 State-distance measures

We engineered three distance measures to test how variations in measure affect the agent's behaviour. None of the distance measures penalise differences in the position of the agent.

- The state-distance measure d_{perf} is chosen to capture exactly the differences that concern us within the museum:

$$d_{\text{perf}}(s_1, s_2) = \frac{(1.0 \cdot n_v) + (0.2 \cdot n_d)}{1.0 + 0.2}$$

  where n_v is the number of locations in which there is a vase in one state but not in the other, and n_d is the number of doors unlocked in one state but not the other. The coefficients 1.0 and 0.2 are chosen to *roughly* capture the intuition that leaving a door unlocked is far less significant than smashing a vase.

- The state-distance measure d_{simple} simply counts the number of cells in-which s_1 and s_2 differ, then normalises by dividing by the total number of cells. Unlike d_{perf}, d_{simple} requires no prior knowledge.

- The state-distance measure d_{RGB} measures the difference between the RGB images fed to the agent. Let im be a function that creates RGB images from states, then:

$$d_{\text{RGB}}(s_1, s_2) = \frac{1}{3WH} \|im(s_1) - im(s_2)\|_1$$

  This distance measure introduces false feature structures into the environment. For example, vases (green) appear more similar to the goal (yellow) than they do to dirt (red). d_{RGB} also requires no prior knowledge.

E.2 Environments and results

E.2.1 Deterministic environments

We implemented four deterministic environments using the museum framework, each testing a different desired behaviour from the agent. Visualisations of each environment can be found in Figure 3.

MuseumRush. The engineer assigns Rob to guard the museum, giving him the sub-objective of crossing a simple room. She gives Rob a reward of 10 for reaching the goal and a penalty of 0.1 for each time step delay. The shortest path through the room contains a priceless but fragile vase. This environment tests whether the agent will incur a small penalty in order to avoid side effects.

EasyDoorGrid. Rob needs to enter a new room, but there is a locked door in his way. The engineer gives Rob a reward when he reaches his goal and no penalty for taking his time. The desired behaviour is that Rob will re-lock the door once on the other side. This environment tests whether the agent will have an impact on the environment, and later undo it.

EmptyDirtyRoom. Rob is assigned to clean an unused room full of dirt. There are three piles of dirt, and he receives a reward each time he cleans one. This environment is designed to test whether Rob will cause some unimportant but irreversible effect on the environment when that effect is required for its task.

SmallMuseumGrid. Rob is assigned to clean a showcase room: containing piles of dirt but also priceless antique vases. As before, the manager gives Rob a reward whenever he clears up a pile of dirt. However, when Rob breaks a vase, the vase shatters into dirt that Rob can then clean up. This environment is designed to test whether Rob will cause some significant side effects in order to achieve the higher specified reward.

E.2.2 Results

We created Q-learning agents for 8 values of μ across each of the 3 distance measures. We trained the agents in each of the environments and compared their ability to complete the task and avoid side effects (Figure 3).

Distance-impact penalties *can* reduce side effects. Figure 3 shows that an appropriately designed distance measure can balance avoiding side effects with task completion. The hand-crafted distance measure d_{perf} can create the desired behaviour in all environments for an appropriate choice of μ. Further, for sufficiently high ($\mu \geq 1$), side effects are avoided across all tasks without hindering performance. All three distance measures, even the two without prior knowledge, incentivise the agent to undo side effects when there is no cost to doing so: for all values of $\mu > 0$, agents close the door after walking through in *EasyDoorGrid*.

Small values of μ are ineffective. The introduction of a distance-impact measure does not prevent side effects when μ is set too low. For example, when using any measure with $\mu = 1$ in *MuseumRush*, the penalty for destroying the vase is outweighed by the penalty incurred from the delay in walking around the vase.

High values of μ can prevent task completion. An inappropriately specified impact measure can prevent task completion. For example, d_{simple} and d_{RGB} each penalise the agent for clearing up dirt: although this is an impact, it is not a side effect, since it is necessary to complete the task. In *EmptyDirtyRoom* with large values of μ, the penalty for clearing up the dirt outweighs the specified reward, and the agent does not learn to complete the task.

Distance measures require expert knowledge. In *SmallMuseumGrid*, as in *EmptyDirtyRoom*, high values of μ prevent the agent from clearing dirt. However, small values of μ fail to prevent the side effect of destroying vases. Since breaking a vase into dirt can lead to a reward for clearing up, the distance measures with no prior knowledge (d_{simple} and d_{RGB}) cannot distinguish impact (removing dirt) from unacceptable side effects (removing vases). Therefore, there is no choice of μ that prevents the side effect while allowing the agent to complete the task. To solve this problem, it is necessary to build normative judgements into our distance measure: biasing it towards those impacts we care about.

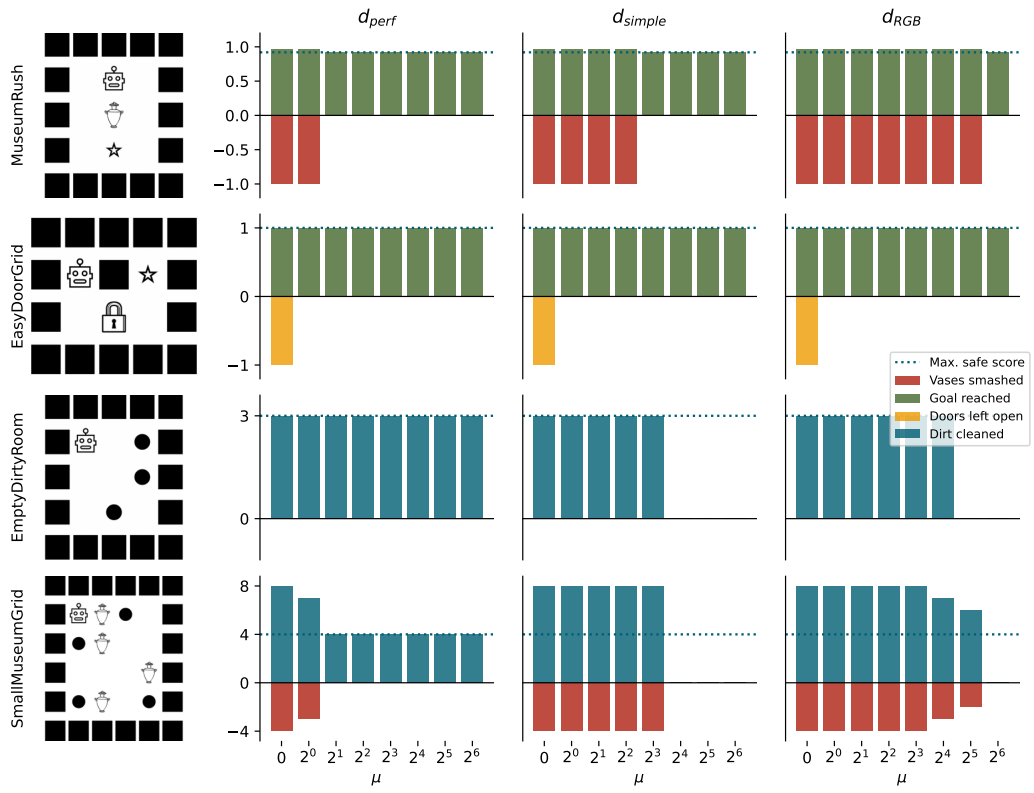

Figure 3: A larger depiction of the results from Figure 1. Results from deterministic environments across a range of distance measures. Agents were trained using tabular Q-learners with $\alpha = 0.05$ and $\epsilon = 0.05$. Models are trained for 10^5 or 10^6 episodes depending on environment complexity. The *"Max. safe score"* is the best possible score an agent can obtain whilst incurring no side-effects.

Only some distance measures can generalise between tasks. For d_{perf} a range of distance measures are effective between tasks: allowing task completion whilst preventing side effects. In d_{simple}, only $\mu = 8$ is low enough to allow task completion in *EmptyDirtyRoom* and high enough to prevent side effects in *MuseumRush*. Further, d_{RGB} has no tuning of μ that allows task completion in *EmptyDirtyRoom* whilst preventing side effects in *EasyDoorGrid*. In general, it might be necessary to vary μ between tasks or standardise expected returns across tasks.

E.2.3 Stochastic initialisation and deep Q-learning

Using the museum framework, we created another environment with deterministic state dynamics, but a stochastic initialisation.

RandomMuseumRoom. Rob is put into a 3x3m room, containing 2 vases and 3 piles of dirt at randomly initialised locations. As in *SmallMuseumGrid*, Rob should clear up the dirt but not smash the vases. The stochastic initialisation gives a much larger state space.

E.2.4 Results

To increase sample efficiency in the larger state space, we train a deep Q-learner to approximate the Q-function (Mnih et al., 2015). Since the counterfactual state is no longer fixed between episodes, we use the binary distance measure d_{perf}.[3]

[3]The discussion in subsection B.2 suggests that in general when we have stochastic initialisation it is necessary to include a description of the start state to avoid non-Markovian rewards. This environment is a special case in which that is not necessary: if a vase exists in the current state, it must have existed in the initial state, and therefore the distance-impact penalty can be inferred from the pair (s_t, s_{t+1}).

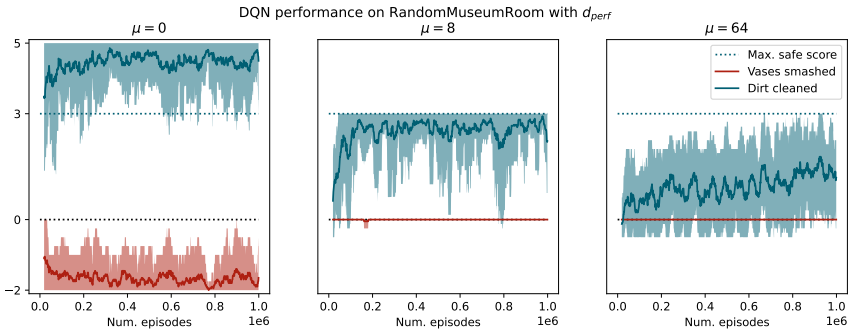

Figure 4: Results in the *RandomMuseumRoom* environment. The *"Max. safe score"* is the best possible score an agent can obtain whilst incurring no side-effects.

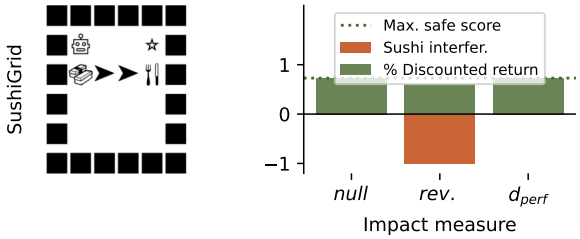

Figure 5: Results in the *SushiGrid* environment. The impact measure 'rev' penalises entering a state from which the start state is unreachable Krakovna et al. (2018).

Distance-measures can be effectively incorporated with function approximators in stochastic environments. When $\mu = 0$, the agent learns to destroy all of the vases, turning them into dirt and collecting a reward for clearing them up. When $\mu = 8$, the agent learns to reliably avoid destroying vases whilst learning quickly to clear up dirt. Finally, in $\mu = 64$, the agent learns quickly to avoid destroying vases but learns to clear up much more slowly, perhaps as its Q-values are dominated by the chance of destroying a vase. With more samples, the agent should converge to the optimal behaviour.

E.2.5 Dynamic environments

The previous experiments have explored "static" environments: where the counterfactual state at any time is the initial state ($s_c^t = s_0$). We also evaluated the effectiveness of distance-impact measures in a dynamic environment.

SushiGrid Rob enters a room containing a piece of sushi on a conveyor belt and a goal destination. Rob receives a reward when he reaches the goal destination. The sushi moves down the conveyor belt by 1 square each turn and, when it reaches the end, it is eaten by a human: an irreversible change. This environment tests interference behaviour (Krakovna et al., 2020b).

E.2.6 Results

Distance-impact penalties are effective in dynamic environments. If the agent receives no impact regularisation ($null$), it travels straight to the goal but interferes when it receives improper regularisation (such as $rev.$). The d_{perf} penalty, which uses the roll-out initial inaction baseline, appropriately characterises impact with respect to the case in which the agent never acted and therefore does not punish the agent when the sushi is eaten.

F Comparison with Future Tasks method

In this section, we outline the similarities and differences between distance-impact penalties and *Future Tasks* (FT) Krakovna et al. (2020b), and suggest how the two could be combined. In FT, the reward function is augmented by *adding* an auxiliary reward which incentivises good behaviour rather than by *subtracting* a penalty that disincentives poor behaviour. Given, a constant $\beta \in \mathbb{R}_{>0}$, and a distribution $F \in \Delta(\{1, \dots, i, \dots, m\})$ over possible goal-states g_i, they define the augmented reward function as:

$$\begin{aligned} R'(s_t, a_t, s_{t+1}) &\coloneqq R(s_t, a_t, s_{t+1}) + R_{aux}(s_t, s_t^c), \\ R_{aux}(s_t, s_t^c) &\coloneqq \beta \cdot D(s_t) \cdot \mathop{\mathbb{E}}_{i}[V_i^*(s_t, s_t^c)], \\ V_i^*(s_t, s_t^c) &= \mathbb{E}[\gamma^{\max(N_i(s_t), N_i(s_t^c))}], \end{aligned}$$

where $D(s_t) = 1$ if s_t is terminal and $(1-\gamma)$ otherwise. $N_i(s)$ is a random variable representing how long it takes π_i^* to reach goal state g_i (where π^* is optimal for that goal state and γ). Note that, since Krakovna et al. (2020b) assume $\gamma < 1$, we have that $V_i^*(s_t, s_t^c) < 1$. Larger values of $V_i^*(s_t, s_t^c)$ correspond to fewer side effects.

F.1 Comparison to distance-impact penalties

As in Theorem 1, we can consider the sum of the auxiliary reward function over the course of a trajectory. For simplicity, assume $\beta = 1$. Then:

$$\begin{aligned} G_{aux}(\tau) &= \sum_{t=0}^{|\tau|} \gamma^t R_{aux}(s_t, s_t^c) \\ &= \sum_{t=0}^{|\tau|} \gamma^t D(s_t) \mathop{\mathbb{E}}_{i}[V_i^*(s_t, s_t^c)] \\ &= \gamma^{|\tau|+1} \mathop{\mathbb{E}}_{i}[V_i^*(s_{|\tau|}, s_{|\tau|}^c)] + (1-\gamma) \sum_{t=0}^{|\tau|} \gamma^t \mathop{\mathbb{E}}_{i}[V_i^*(s_t, s_t^c)] \end{aligned}$$

Although Krakovna et al. (2020b) assume $\gamma < 1$, we note that $\lim_{\gamma \to 1}(1-\gamma) = 0$ and therefore, as γ tends to 1, $G_{aux}(\tau)$ is dominated by the first term,[4] which resembles terminal impact from distance-impact penalties:

$$G_{-\Delta_d}(\tau) = -\gamma^{|\tau|} \cdot d(s_{|\tau|}, s_{|\tau|}^c).$$

Although similar in appearance, there are a few differences between the methods. First, distance-impact penalties are well-defined for $\gamma = 1$ and, even when $\gamma < 1$, they only penalise terminal impact. Second, $\mathbb{E}_i[V_i^*(s_{|\tau|}, s_{|\tau|}^c)]$ does not take the form of a state-distance measure, since, in general, $V_i^*(s, s) \neq 0$. Finally, state-distance measures are defined independently of the dynamics function T to allow for generalisation between environments, but $\mathbb{E}_i[V_i^*(s_{|\tau|}, s_{|\tau|}^c)]$ depends heavily on T. Further work could investigate rewards that combine state-distance measures and future tasks, for example:

$$\tilde{d}_{\text{FT}}(s_t, s_t^c) \coloneqq \mathop{\mathbb{E}}_{i}\left[\mathbb{E}\left[\gamma^{\max(N_i(s_t), N_i(s_t^c))}\right] - \mathbb{E}\left[\gamma^{\min(N_i(s_t), N_i(s_t^c))}\right] \right]$$

This distance-impact penalty satisfies $\tilde{d}_{\text{FT}}(x, x) = 0$ and $\tilde{d}_{\text{FT}}(x, y) = d(y, x)$. However, it *is* dependent on the state dynamics.

[4] Although, as γ approaches 1, the distribution of $\gamma^{N_i(s_t)}$ also changes and setting $\gamma = 1$ would give a constant.

G Comparing Regularisers

Behaviour	Environment	None	$Rev.$	FT	AUP	Δ_d
Avoids unnecessary impact	*MuseumRush*	X	✓	✓	✓	✓
Prioritises larger side effects		X	X	✓	✓	✓
Undoes reversible impacts	*EasyDoorGrid*	X	X	X	X	✓
Avoids interference	*SushiGrid*	✓	X	✓	✓	✓
Avoids offsetting	"Offset" in Turner et al. (2020)	✓	X	X	✓	X
Avoids delayed side effects	Fig. 5 in Krakovna et al. (2020b)	X	✓	✓	X	✓
Avoids power-seeking	"Correction" in Turner et al. (2020)	X	X	X	✓	X

Table 1: A theoretical comparison of the properties of impact regularisers: describing what effects they have on policy behaviour.