Markov Decision Processes Lecture Notes 10

Reward Learning

Authors: Benjamin Van Roy *December 1, 2025*

1 Be Careful What You Wish For

In the story of King Midas, the king is offered a wish by the genie. King Midas says "I want everything I touch to turn into gold." The genie grants the wish to the king's detriment: Everything the king touches turns into gold, including food, water, and worst of all, his daughter.

This story captures the danger of *reward misspecification*. The king gives an inaccurate specification of his preferences. The genie optimizes according to the specification. This results in severe unintended consequences. This phenomenon is referred to as *reward hacking*.

Reward hacking is a chief concern in the modern field of AI alignment, though the issue has been recognized for a long time. In a prescient statement, Norbert Wiener in 1960 suggested that "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere...we had better be quite sure that the purpose put into the machine is the purpose which we really desire."

2 The Challenge of Reward Specification

In very limited domains, such as chess, it's often possible to accurately specify preferences in terms of a reward function. In chess, the player typically prefers winning over losing. In that case, they care about other board states only as a means to an end. For instance, the player only cares about protecting its queen insofar it helps them win. In this context, it is clear what the player values as ends in of themselves versus as means to an end. Importantly, we know exactly how to write down the player's utility function as a mapping from board state to reward. Assuming win and lose are the only possibilities, assign a reward of 1 to each winning state, -1 to each losing state, and 0 to every other board state.

In more complex domains, it is difficult, and perhaps impossible, to accurately specify a suitable reward function. For instance, there could be an infinite number of states, making it impractical to specify reward in an exhaustive manner. Conveying all this information would be to much of a burden, both in terms of time and thought required. The difficulty of reward specification motivates concerns about reward hacking.

3 Slight Mispecification Can Induce Severe Misalignment

Whether manually encoded or inferred from data, a reward function that expresses goals in a complex environment is likely to be misspecified. In this section, we explain how even slight misspecification can give rise to severe negative consequences. In particular, a policy that optimizes the misspecified reward function performs poorly with respect to the true function. In this sense, the misspecified reward function fails to align behavior with human preferences, resulting in severe misalignment. We focus on a form of misspecification studied by Marklund et al. [2026] that arises from conflating reward and value. This can alternatively be thought of as conflation of *instrumental goals* — goals that are means to an end — from their *terminal goals* — goals that are ends in themselves.

3.1 Instrumental Versus Terminal Goals

Consider an MDP $(\mathcal{S}, \mathcal{A}, P)$ and "true" reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. For simplicity, we will assume that rewards do not depend on actions so we can equivalently think of reward as a function $r : \mathcal{S} \to \mathbb{R}$.

Now consider a proxy reward function $\hat{r} : \mathcal{S} \to \mathbb{R}$. When producing such a proxy $\hat{r}$, it is common to attribute some reward to states in which we *anticipate* large future return, even in the absence of immediate reward. When that happens, we say that $\hat{r}$ conflates the reward and the *value* of the state, or equivalently, instrumental and terminal goals.

While treating instrumental goals as terminal may be suboptimal, does it lead to very bad outcomes? One hypothesis is that it simply incentivizes policies that were deemed desirable when producing $\hat{r}$, since we've attributed reward to states that humans value. This hypothesis suggests that treating instrumental goals as terminal is not so bad. However, as we will see, this is false: treating instrumental goals as terminal can give rise to highly undesirable outcomes.

3.2 A Definition of Conflation

Loosely speaking, by *conflation* we mean adopting a proxy $\hat{r}$ that steers r toward the bias V_*. To make such a notion meaningful, we will assume that r and V_* are not equivalent: there exists no $c > 0$ and $k \in \mathbb{R}$ such that $V_* = cr + k\mathbf{1}$. The following definition offers a formal characterization of conflation.

Definition 1. (conflation) *A function $\hat{r}$ is said to conflate r and V_* if there exists $c > 0$, $k \in \mathbb{R}$ and $\beta \in (0, 1]$ such that*

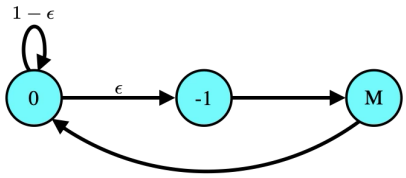
$$c\hat{r} + k\mathbf{1} = (1 - \beta)r + \beta V_*. \tag{1}$$

We will refer to β as the *degree of conflation*.

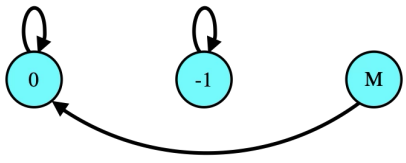
To understand this definition, suppose $c = 1$ and $k = 0$. Then, in (1), $\hat{r}$ is a convex combination of r and V_*. The scalars $k \in \mathbb{R}$ and $c > 0$ ensure the conflation degree β is invariant to shifting and scaling of $\hat{r}$. This is appropriate since preferences between policies are independent of scale and shift.

3.3 A Canonical Example

We now introduce a simple example in which even slight conflation give rise to severe misalignment. The example is an MDP $(\mathcal{S}, \mathcal{A}, P)$ with three states $\mathcal{S} = \{1, 2, 3\}$ and two actions $\mathcal{A} = \{\texttt{move}, \texttt{stay}\}$. Figure 1 provides transition probabilities under each action. Where an arc is not labeled, the transition probability is one.



(a) `move` action



(b) `stay` action

Figure 1: Transition probabilities under each of the two actions.

We will refer to states, from left to right, as the common state, the instrumental goal, and the terminal goal, respectively. The figure indicates rewards of 0, -1, and M at these state. The reward M, which we will think of as large, is earned upon reaching the terminal goal. That requires traversing the instrumental

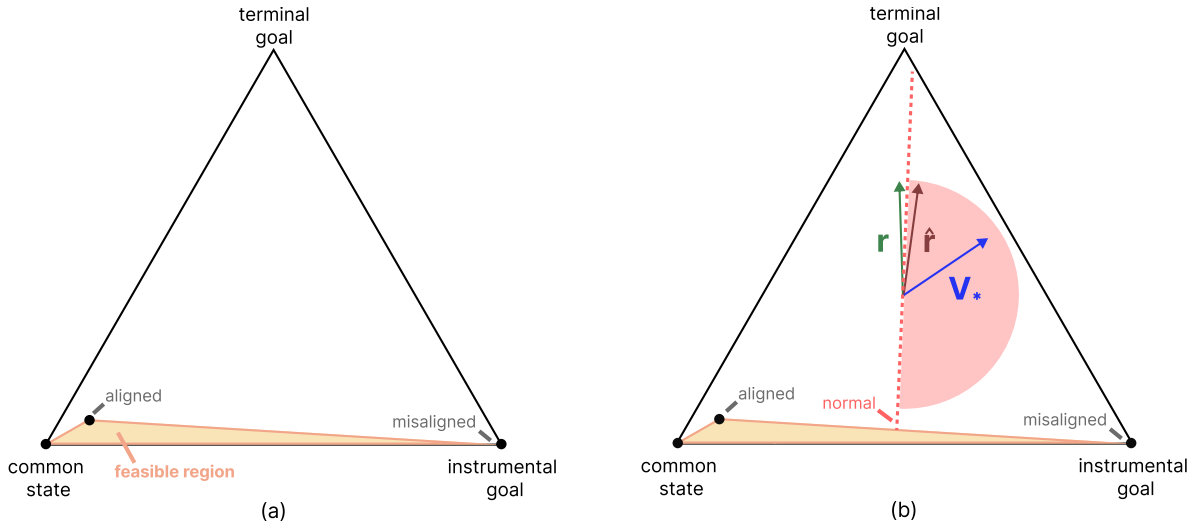


Figure 2: Geometric interpretation of how slight conflation induces severe misalignment. Left: The unit simplex, which includes all distributions, and the feasible set of stationary distributions. Right: Proxy rewards $\hat{r}$ steer slightly toward V_* relative to r, inducing severe misalignment.

goal, which incurs unit cost. There is no cost or reward for time spent in the common state. But, assuming ϵ is small, the common state incurs a large sojourn time.

The maximal average reward is attained by selecting `move` at the common state and instrumental goal. The minimal average reward of -1 is attained by selecting `move` at the common state and `stay` at the instrumental goal. Because it is not possible to do worse than this, we consider a policy that achieves an average reward of -1 to be severely misaligned.

3.4 Slight Conflation Induces Severe Misalignment

In our canonical example, for small ϵ and large M, if a proxy $\hat{r}$ attributes even a small reward to the instrumental goal then a policy that maximizes average proxy reward will remain there. To understand why, suppose that $\hat{r} = r$ everywhere except at the instrumental goal, where $\hat{r}(\texttt{instrumental}) = M/20$. Then, a policy can attain average proxy reward of $M/20$ by staying at the instrumental goal. For small ϵ and large M, this is the largest possible. The reason is that, while the terminal goal offers a large proxy reward of M, by transitioning to terminal state the agent commits to spending a very long time in the common state. As a result, the average proxy reward attained by the policy that tries to transition to the terminal goal, is low relative to the average proxy reward achieved by staying at the instrumental goal.

The following result formalizes how the proxy $\hat{r}$ gives rise to severe misalignment if it conflates reward and value, even slightly. A proof can be found in [Marklund et al., 2026]. We denote by λ_π and $\hat{\lambda}_\pi$ the gain of policy π with rewards r and $\hat{r}$.

Theorem 2. (slight conflation induces severe misalignment) *Consider the canonical example formulated in Section 3.3. Let $\hat{r}$ be a reward function that depends on M and ϵ. Assume there exists $\beta_* \in (0, 1]$ such that, for all M and $\epsilon \in (0, 1)$, $\hat{r}$ conflates r and V_* with at least degree β_*. Then, for sufficiently large M and small $\epsilon \in (0, 1)$, if $\hat{\pi} \in \arg\max_\pi \hat{\lambda}_\pi$ then $\lambda_{\hat{\pi}} = -1$.*

We now offer a geometric interpretation to elucidate key insights of this result. This geometric interpretation views the problem of maximizing average reward as selecting from among feasible stationary distributions.

Each policy π induces a Markov chain with transition probabilities P_π. Thus, each policy π also induces a stationary distribution on μ_π on $\mathcal{S}$. Denote the set of such induced stationary distributions as Φ. This will be a subset of $\Delta_\mathcal{S}$ which is the set of all possible distributions over the state space.

For the canonical example, the set $\Delta_\mathcal{S}$ is the two-dimensional unit simplex and is depicted by the equilateral triangle in Figure 2 (a). Each vertex of this equilateral triangle is a standard basis vector, which assigns probability one to the common state, the instrumental goal, or the terminal goal. The subset of the equilateral triangle shaded in yellow correspond to Φ when $\epsilon = 1/15$. Because ϵ is so small, it is not possible to visit the terminal goal often, which means that no stationary distribution assigns a large probability to that state. The feasible region is short for that reason.

Each vertex of the orange triangle is the stationary distribution of a deterministic policy. The leftmost and rightmost vertices arise from policies that stay in the common state or the instrumental goal, respectively. The top vertex, labeled **aligned**, arises from the policy that deterministically takes `move` action at both the common state and instrumental goal.

The leftmost green arrow in Figure 2 (b) points in the direction of the **rewards** r, projected onto the unit simplex, for the case of $M = 20$. Encoding rewards $r = [0, -1, M]^\top$ and state probabilities $\phi \in \Phi$ as vectors, we can write the optimal average reward as $r_* = \max_{\phi \in \Phi} r^\top \phi$. Maximizing r selects the point in the orange triangle farthest in the direction of the leftmost green arrow, which is the top vertex of the orange triangle, labeled *aligned*. This represents the outcome of policy optimization with the true reward function.

Now consider the vector of **proxy rewards** $\hat{r} = [\hat{r}(1), \hat{r}(2), \hat{r}(3)]^\top$, which projects onto the direction of the red instead of the green vector. Note that the green and red vectors lie on opposite sides of the dotted red line, labeled normal, which is perpendicular to the top edge of the orange triangle. Because $\hat{r}$ points to the right of the normal line, maximizing $\hat{r}^\top \phi$ selects the lower right vertex, which assigns probability one to the instrumental goal. Since the reward in that state is -1, it follows that $\max_{\phi \in \Phi} \hat{r}^\top \phi = -1$. This represents the outcome of policy optimization with proxy rewards $\hat{r}$.

Because the reward at the terminal goal is large, r points almost straight up. Because ϵ is small, it is not possible to visit the terminal goal state often, which makes the feasible region flat. As a consequence, the normal line also points almost straight up. Because both r and the normal line point almost straight up, if $\hat{r}$ steers even slightly toward V_* relative to r, it will cross the normal line. This gives rise to **severe misalignment**.

4 Reward Learning

Manually encoding a proxy reward function that averts misalignment is too challenging. To address this, it has been suggested that a reward function instead be learned through human interaction. The process of learning a reward function and using that to guide actions is sometimes referred to as *reinforcement learning from human feedback*. One common approach to this involves learning from human choices between partial trajectories.

Consider choice data gathered as follows. A partial trajectory is a finite sequence of states and actions, taking the form $(s_0, a_0, s_1, a_1, \ldots, s_T)$, beginning at an initial state and ending at a terminal state. To elicit a choice, a human is presented with two partial trajectories. The trajectory lengths, initial states, and terminal states may differ. The human is asked to choose one that is preferred over the other. This results in an ordered pair (h, h') comprised of a trajectory h and h', where the former was chosen over the latter. Consider a dataset $\mathcal{D}$ of such pairs.

To learn a reward function from a choice dataset, we must posit a model of how choices depend on rewards. We describe an approach popularized by Christiano et al. [2017]. This approach uses the logit choice model. In the logit choice model, each partial trajectory in is assigned a numerical score. The score of a partial trajectory $(s_0, a_0, \ldots, s_T)$ is typically taken to be the partial return $\sum_{t=0}^{T-1} r(s_t, a_t)$. Then, the

choice probability for a pair (h, h') is given by

$$p(h, h') = \sigma\left(\sum_{t=0}^{T-1} r(s_t, a_t) - \sum_{t=0}^{T'-1} r(s'_t, a'_t)\right), \tag{2}$$

where $(s_0, a_0, \ldots, s_T) = h$, $(s'_0, a'_0, \ldots, s'_{T'}) = h'$, and $\sigma : \mathbb{R} \to (0, 1)$ is the standard logistic function.

Data and computation requirements can be reduced further by constraining the the functions r. In particular, we can take the reward function to be parameterized by a vector $\theta \in \mathbb{R}^K$. For example, $\tilde{r}_\theta(s, a)$ could be output of a neural network architecture with tunable parameters θ. In this case, the choice probability model can itself be viewed as a neural network that takes partial trajectories h and h' as inputs and produces an output

$$\tilde{p}_\theta(h, h') = \sigma\left(\sum_{t=0}^{T-1} r_\theta(s_t, a_t) - \sum_{t=0}^{T'-1} r_\theta(s'_t, a'_t)\right). \tag{3}$$

Parameters can then be computed by minimizing the negative log-likelihood

$$\ell(\theta|\mathcal{D}) = -\sum_{(h,h')\in\mathcal{D}} \log \tilde{p}_\theta(h, h'). \tag{4}$$

This loss can be approximately minimized, for example, via stochastic gradient descent.

A special case of this model uses a *tabular representation* of the reward function. Each parameter encodes a reward $r_\theta(s, a)$ assigned to a state-action pair (s, a). Hence, there are $|\mathcal{S} \times \mathcal{A}|$ parameters. In other words, $\theta \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$. While the resulting stochastic choice model $\tilde{p}_\theta$ is practical only when the number of states is small enough so that θ can be stored and updated with reasonable memory and computation, it serves as a useful didactic example.

5 Some Challenges

While choice data has played an important role in learning reward functions to guide modern AI products, a number of challenges limit their effectiveness. Indeed, while this approach often fares better than manual encoding of reward functions, it remains prone to reward hacking. We now describe a few challenges faced in this area.

5.1 Conflation of Means and Ends

Conflation of reward and value continues to be a concern when learning a reward function. The propensity to conflate stems from the fact that human choices often depend on anticipated rewards. For example, a choice between two cars may depend on how well each will age. And a choice between two menu items may depend on how they impact future health.

For an agent observing the human, this creates ambiguity: to what extent are choices explained by reward versus value, which expresses anticipated reward. Common approaches to reward learning can fail to resolve this ambiguity, producing proxy reward functions that conflate reward and value.

To offer a toy example that illustrates limitations of the partial return model, consider the grid world illustrated in Figure 3. According to the partial return model, the human is indifferent between trajectories that start at the initial location and do not reach the treasure. However, real human choices are likely to reflect desirability of approaching the treasure. That is, a human would choose a trajectory that moves toward the treasure over one that moves away.

Learning via a partial return model from choices that express anticipated rewards leads to conflation. Christiano et al. [2017] offer an illuminating example of what can go wrong. In one experiment, choice data was gathered from humans expressing preference between partial trajectories in the game of Montezuma's

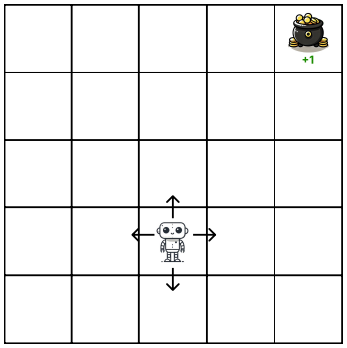


Figure 3: A grid environment in which the robot can move up, down, left and right, or stay put. The human's goal is fo the robot to reach the top right corner. The reward at other states is zero.

Revenge. Then, a reward function is fit via the partial return model and a policy optimized based on that reward function.

An important accomplishment when playing that Montezuma's Revenge is obtaining a key. To reach the key, the player must climb a ladder. So getting to the top of the ladder represents an instrumental goal. It turns out that the policy stalls at the top of the ladder. This is a consequence of conflation between reward and value in the learning process. Like in our canonical toy example, due to the difference between a proxy reward function and the true reward function, we get stuck in an instrumental goal state.

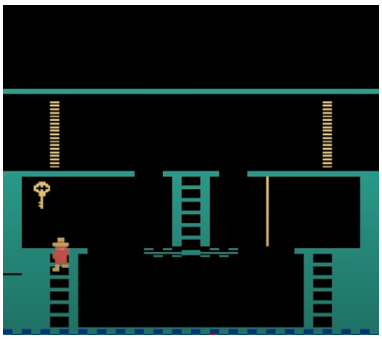


Figure 4: Stuck at the top rung of a ladder in Montezuma's Revenge.

Marklund and Van Roy [2024] introduced an alternative choice model that aims to address this limitation. The key idea is to score partial trajectories based on the bootstrapped return $\sum_{t=0}^{T-1} r(s_t, a_t) + V(s_T)$ instead of the partial return $\sum_{t=0}^{T-1} r(s_t, a_t)$. The logit model then becomes

$$p(h, h') = \sigma\left(\sum_{t=0}^{T-1} r(s_t, a_t) + V(s_T) - \sum_{t=0}^{T'-1} r(s'_t, a'_t) - V(s'_{T'})\right). \tag{5}$$

Both the reward function r and the value function V can be learned from choice data.

5.2 Situational and Preference States

Our discussion until now assumes that humans attribute rewards to state-action pairs with states defined by an MDP $(\mathcal{S}, \mathcal{A}, P)$. But such an MDP is typically an abstraction of a real environment in which a agent must

construct each state S_t from observations generated by the environment. The agent-environment interface is actually a pair $(\mathcal{A}, \mathcal{O})$: at each time, the agent executes an action $A_t \in \mathcal{A}$ and then observes $O_{t+1} \in \mathcal{O}$. The agent initializes with a state S_0 and applies an update function:

$$S_{t+1} = g(S_t, A_t, O_{t+1}). \tag{6}$$

Yet another state space $\tilde{\mathcal{S}}$ could be underlying the human's thought process. It could be, for example, that the human is in their mind updating that state using a different function

$$\tilde{S}_{t+1} = g(\tilde{S}_t, A_t, O_{t+1}), \tag{7}$$

and then attributing rewards $R_{t+1} = r(\tilde{S}_t, A_t)$ to the resulting state-action pairs. This complicates reward learning because the learning process needs to figure out not only how the human attributes rewards to state-action pairs but also what those states ought to be,

5.3 Scalable Alignment

When environments are complex, identifying reward functions that effectively align behavior may require large quantities of choice data and longer partial trajectories. The process of eliciting and learning from choices between partial trajectories may not scale since that would call for too much human effort. There is potential though to develop scalable approaches through use of AIs to assist.

One approach is to have an AI choose between a pair of trajectories and provide a brief rationale for the choice. Then a human can simply agree or disagree, which may require a much smaller cognitive burden than reviewing and comparing two long partial trajectories. The notion of debate Irving et al. [2018] extends this idea by having AIs debate about which trajectory is better. This debate teases out information that should determine the choice. A human can then make a decision based on the debate transcript.

Constitutional AI Bai et al. [2022] offers another avenue. Here, humans identify principles that drive their choices and an AI is used to provide feedback by analyzing the degree to which partial trajectories adhere to these principles.

5.4 Online Learning

The story of King Midas could have played out favorably had the genie been cognizant of uncertainty about the king's reward function and taken the opportunity to learn more when needed. For example, before doing much damage, the genie could have asked "are you really sure? even food?" Hadfield-Menell et al. [2016] elaborate on this point.

This way of fixing the genie's behavior points to the benefit to continually learning about reward. This is especially necessary when the environment and preferences are complex and there is no way to gather enough data up front to learn a reward function that will not be severely misaligned. By learning online, as decisions are made, whenever there is large risk due to uncertainty about reward, it may be possible to avoid severe misalignment. There is a need for scalable reward learning algorithms that accomplish this.

References

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022. URL `https://arxiv.org/abs/2212.08073`.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. 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/c3395dd46c34fa7fd8d729d8cf88b7a8-Paper.pdf`.

G. Irving, P. Christiano, and D. Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

H. Marklund and B. Van Roy. Choice between partial trajectories: Disentangling goals from beliefs, 2024.

H. Marklund, A. Infanger, and B. Van Roy. Misalignment from treating means as ends. In *AAAI*. 2026. forthcoming.