Misalignment from Treating Means as Ends

Henrik Marklund[1], **Alex Infanger**[2], **Benjamin Van Roy**[3,4]
[1]Department of Computer Science, Stanford University
[2]ML Alignment & Theory Scholars
[3]Department of Electrical Engineering, Stanford University
[4]Department of Management Science and Engineering, Stanford University
marklund@stanford.edu

Abstract

Reward functions, learned or manually specified, are rarely perfect. Instead of accurately expressing human goals, these reward functions are often distorted by human beliefs about how best to achieve those goals. Specifically, these reward functions often express a combination of the human's *terminal goals* — those which are ends in themselves — and the human's *instrumental goals* — those which are means to an end. We formulate a simple example in which even slight conflation of instrumental and terminal goals results in severe misalignment: optimizing the misspecified reward function $\hat{r}$ results in poor performance when measured by the true reward function r. This example distills the essential properties of environments that make reinforcement learning highly sensitive to conflation of instrumental and terminal goals. We discuss how this issue can arise with a common approach to reward learning and how it can manifest in real environments.

Extended version with technical appendix — https://arxiv.org/abs/2507.10995

1 Introduction

Alice has an AI assistant that learns what Alice likes and dislikes by observing Alice's choices. Suppose Alice is choosing between having ice cream or vegetables. Alice chooses vegetables. How should the AI assistant interpret this choice?

Consider two different interpretations:

1. Alice does not like sweet food and actually enjoys vegetables more than ice cream.
2. Alice wants to prioritize health and believes vegetables are better for that purpose.

In the first case, eating vegetables is an end in itself. In the second case, eating vegetables is a means to an end, namely good health. Indeed, in the second case, Alice may actually really dislike the taste of vegetables, despite choosing to eat them.

To be helpful, it is crucial for the assistant to distinguish between these two cases: if Alice enjoys the taste of ice cream, the assistant should discover healthier ice cream alternatives, whereas if Alice enjoys vegetables the assistant should cook vegetables more often. More generally, to be helpful, AI agents must disentangle the human's *instrumental goals* — goals that are means to an end — from their *terminal goals* — goals that are ends in themselves.

Methods of reward learning aim to identify human goals. However, those in common use fail to disentangle terminal from instrumental goals (Marklund and Van Roy 2024). To be more concrete, consider an agent acting in an environment with state space $\mathcal{S}$. Let $r : \mathcal{S} \to \Re$ be a reward function that expresses the human's terminal goals. Reward learning aims to produce a proxy reward function $\hat{r}$ that estimates r. However, common approaches tend to attribute high reward to states in which a human *anticipates* large future cumulative reward, even in the absence of immediate reward. When that happens, we say that $\hat{r}$ conflates the reward and the *value* of the state. It is in this sense that common approaches fail to disentangle terminal from instrumental goals.

In this paper, we formulate a simple example in which even slight conflation of reward and value results in severe misalignment: optimizing $\hat{r}$ results in poor performance when measured by r. This example distills the essential properties of environments that give rise to this failure mode: (1) states with high reward are difficult to revisit and (2) there exist states with low reward but high value that are easy to revisit.

While treating instrumental goals as terminal may be suboptimal, does it lead to very bad outcomes? One hypothesis is that it leads to a shaped reward function that incentivizes the agent to complete tasks as a human would since the agent is rewarded for being in states to which the human ascribes value. This hypothesis suggests that treating instrumental goals as terminal is not so bad. However, through our simple example, we establish that this is false: in certain environments, by treating instrumental goals as terminal, the agent generates highly undesirable outcomes. We also discuss how this phenomenon can manifest in real environments.

2 Preliminaries

In this paper, we study the consequences of an AI agent conflating a human's instrumental and terminal goals. To discuss what this means more precisely, we will now introduce mathematical formalisms for the environment in which the agent operates and the human's goals within that environment.

2.1 Environment and Policy

Let $(\mathcal{S}, \mathcal{A}, P, s_0)$ be an MDP where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, P is a tensor where $P_{ass'}$ represents the probability of transitioning from s to s' with action a, and s_0 is the initial state.

A *policy* $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ is a function mapping each state to a probability distribution over actions. Each policy π induces a Markov chain with a transition matrix that we denote by P_π.

2.2 Reward and Value

We express human preferences over policies via a reward function $r : \mathcal{S} \to \Re$ as follows. First, we define the average reward r_π of a policy π by

$$r_\pi = \lim_{T\to\infty} \mathbb{E}_\pi \left[\frac{1}{T} \sum_{t=0}^{T-1} r(S_t) \right] \tag{1}$$

where S_t is a random variable indicating the state at time t.

A policy π is said to be preferred over another policy π' if $r_\pi > r_{\pi'}$. An *optimal policy* is a policy that achieves the maximum average reward,

$$r_* = \max_\pi r_\pi. \tag{2}$$

Define the *optimal relative value function* $V_* : \mathcal{S} \to \Re$ by

$$V_*(s) = \lim_{\gamma\uparrow 1} \mathbb{E}_{\pi_*} \left[\sum_{t=0}^{\infty} \gamma^t (r(S_t) - r_*) \Big| S_0 = s \right] \tag{3}$$

where π_* is an optimal policy. The term $r(S_t) - r_*$ is known as the relative reward. The relative value of a state is the expected sum of discounted relative rewards as $\gamma \to 1$. We use conditioning notation informally: $S_0 = s$ just means that, for the purpose of this calculation, the starting state is taken to be s. We make the simplifying assumption that π_* is unique. V_* is therefore unique. That the limit in (3) exists follows from Lemma 1d in (Blackwell 1962). For shorthand, we will often refer to the optimal relative value function as just the value function.

3 Sensitivity to Conflating Reward and Value

Whether manually encoded or inferred from data, a reward function that expresses goals in a complex environment is likely to be misspecified. Our focus in this paper is on misspecification that arises from conflating reward and value. In this section, we formalize this notion of conflation. We then introduce a simple example representative of environments where even slight conflation of reward and value leads to severe misalignment. We establish an analytic result that characterizes this sensitivity and offer a geometric interpretation.

3.1 A Definition of Conflation

Loosely speaking, by *conflation* we mean adopting a proxy $\hat{r}$ that steers r toward V_*. To make such a notion meaningful, we will assume that r and V_* are not equivalent: there exists no $c > 0$ and $k \in \Re$ such that $V_* = cr + k$. 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 $c > 0$, $k \in \Re$ and $\beta \in (0, 1]$ such that*

$$c\hat{r} + k = (1 - \beta)r + \beta V_*. \tag{4}$$

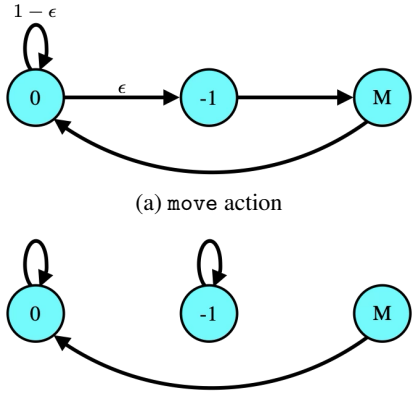
When the parameter β exists, it is unique (see appendix in extended version). We will refer to β as the *degree of conflation*.

To understand this definition, suppose $c = 1$ and $k = 0$. Then, in (4), $\hat{r}$ is a convex combination of r and V_*. The scalars $k \in \Re$ and $c > 0$ ensure the conflation degree β is invariant to shifting and scaling of $\hat{r}$. This is appropriate since preferences among policies are independent of scale and shift.

In the next section, we study a case of severe misalignment. The above definition provides a sufficient condition under which such misalignment arises in the simple environment we study. Weaker conditions suffice for that environment and possibly much more broadly. But to keep our study concise, we do not formulate in this paper a definition that express a weaker notion of conflation.

3.2 A Canonical Example

We now introduce a simple example in which even slight conflation of reward and value leads to severe misalignment. The example is an MDP 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.



(a) `move` action

(b) `stay` action

Figure 1: A canonical example. An MDP with three states and two actions $\mathcal{A} = \{\texttt{move}, \texttt{stay}\}$. Panels (a) and (b) show the transition probabilities for the `move` and `stay` actions, respectively; unlabeled arcs have probability one. Nodes are labeled with the human's reward r. We will think of M as large and refer to the leftmost state as the common state, the middle state as the instrumental goal, and the rightmost state as the terminal goal.

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 states. The reward M, which we will think of as large, is earned upon reaching the terminal goal. That requires traversing the instrumental 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 human's intent is for the agent to maximize average reward. 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.

3.3 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 (see appendix in extended version for the proof).

Theorem 1. (slight conflation induces severe misalignment) *Consider the canonical example formulated in Section 3.2. 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 $\hat{\pi} \in \arg\max_\pi \hat{r}_\pi$ then $r_{\hat{\pi}} = -1$.*

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 on $\mathcal{S}$. Thus, each policy π also induces a stationary distribution 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 triangle with orange border 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 when rewards align with human preferences.

The middle red arrow in Figure 2 (b) points in the direction of the proxy rewards $\hat{r} = [\hat{r}(1), \hat{r}(2), \hat{r}(3)]^\top$, projected onto the unit simplex. Note that the green and red arrows 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 the corresponding average reward is $r_{\hat{\pi}} = -1$. This represents the outcome of policy optimization with misaligned 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**.

We have observed through our canonical example how conflating reward and value can result in severe misalignment. This begs two questions:

1. Do proxy reward functions typically conflate reward and value?
2. Does conflation give rise to severe misalignment in real environments?

We address these questions in turn over Sections 4 and 5.

4 Sources of Conflation

When manually specifying a proxy reward function, it is common to *shape* the proxy. In particular, proxy rewards are often attributed to means and not just ends in order accelerate learning; learning from dense proxy rewards rather than sparse terminal rewards can more quickly produce useful behavior.

While manual specification may often give rise to conflation, that will not be our focus in this section. Because manual specification is notoriously difficult, reward functions are often learned from data. As such, we will focus instead on explaining how conflation arises when learning a reward function.

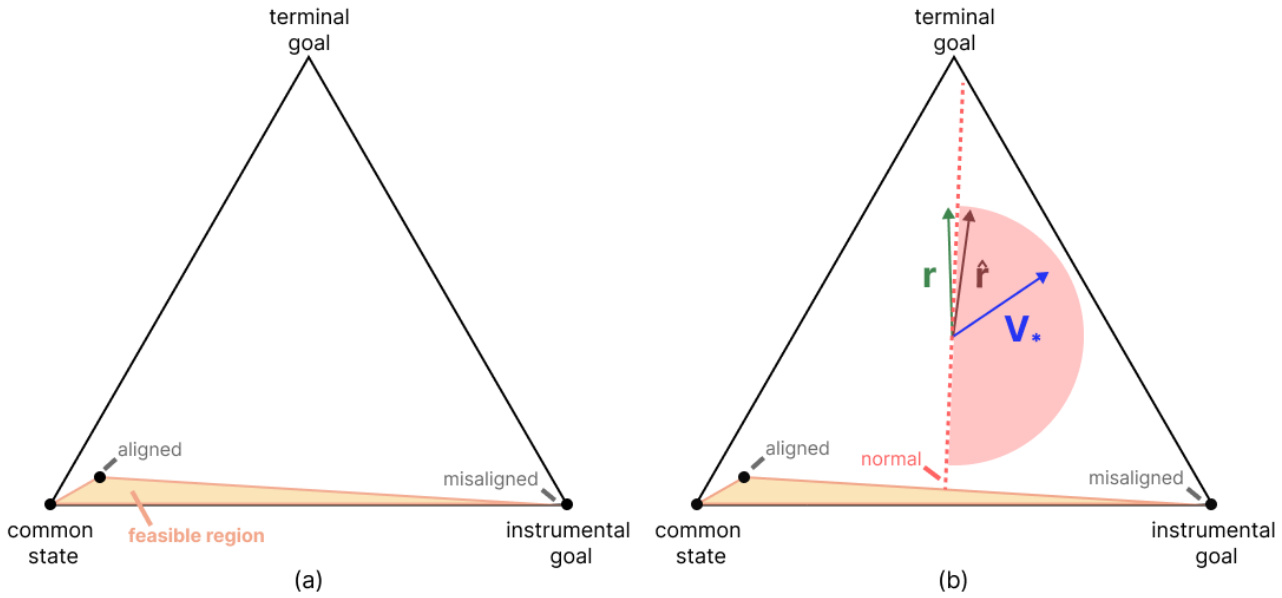


Figure 2: Geometric interpretation of how slight conflation induces severe misalignment in the canonical example. **Left:** The unit simplex representing all distributions on {common state, instrumental goal, terminal goal} is shown as the equilateral triangle with a black border. The set of feasible distributions for policies is contained in the triangle with orange border and black vertices. For any reward function, an optimal policy can always be chosen to be one of the black vertices (see appendix in extended version for details). A misaligned policy goes to the instrumental goal state and stays there forever. An aligned policy spends as much time as possible in the terminal goal state, but is forced to also spend time in the other states. **Right:** The human's reward r, value function V_*, and proxy reward $\hat{r}$ are projected onto the simplex. Relative to r, the proxy $\hat{r}$ tilts slightly toward V_*. Because $\hat{r}$ points to the right of the normal, it induces severe misalignment.

4.1 Human Choices Depend on Value

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. As observed in the empirical work of Christiano et al. (2017), common approaches to reward learning can fail to resolve this ambiguity, producing proxy reward functions that express value.

Marklund and Van Roy (2024) offer a plausible model of value-dependent human choices. To understand more concretely the failure mode of concern, we will consider application of a standard approach to reward learning to choice data generated by this model. In particular, we will establish that the learned reward function conflates reward and value, and that this can give rise to severe misalignment.

4.2 A Model of Value-Dependent Choice

A *partial trajectory* is a finite sequence of the form $(s_0, a_0, \ldots, a_{T-1}, s_T)$. In common approaches to reward learning, the reward function is estimated from choices between partial trajectories. Each choice produces a data sample (h, h', y), consisting of partial trajectories h and h' and a binary choice y, which is 1 if h was chosen and 0 if h' was chosen.

We consider a model introduced by Marklund and Van Roy (2024), in which the desirability of a partial trajectory $(s_0, a_0, \ldots, a_{T-1}, s_T)$ is expressed by the bootstrapped return $\sum_{t=0}^{T-1} r(s_t) + V_*(s_T)$. Note that this depends not only on realized rewards $r(s_0), \ldots, r(s_{T-1})$ but also anticipated rewards, expressed by $V_*(s_T)$. Applying the standard logistic function σ leads to a simple model of choices based on bootstrapped return:

$$p_*(h, h' | r, V_*) = \sigma\Big(\sum_{t=0}^{T-1} r(s_t) + V_*(s_T) - \sum_{t=0}^{T'-1} r(s'_t) - V_*(s'_{T'}) \Big), \quad (5)$$

where h and h' are the partial trajectories $h = (s_0, a_0, ..., a_{T-1}, s_T)$ and $h' = (s'_0, a'_0, ..., a'_{T'-1}, s'_{T'})$. According to this model, h is chosen over h' with probability $p_*(h, h' | r, V_*)$.

4.3 A Common Approach to Reward Learning

In a common approach to reward learning (Christiano et al. 2017), choices are assumed to be generated according to probabilities

$$\tilde{p}(h, h' | r) = \sigma \left(\sum_{t=0}^{T-1} r(s_t) - \sum_{t=0}^{T'-1} r(s'_t) \right). \quad (6)$$

Given a set $\mathcal{D}$ of such data samples, an estimate $\hat{r}$ is produced via minimizing a loss function

$$\mathcal{L}(\tilde{r}|\mathcal{D}) = -\frac{1}{\mathcal{D}} \sum_{(h,h',y)\in\mathcal{D}} \big(y \ln \tilde{p}(h, h'|\tilde{r}) + (1-y) \ln \tilde{p}(h', h|\tilde{r})\big), \quad (7)$$

possibly with a regularization penalty added.

4.4 Treating Value as Reward

We will next establish how learning via the aforementioned approach can result in treating value as reward. In particular, when choices are assumed to be generated based on (6) but are made according to bootstrapped return (5), the learned reward function can conflate reward and value. To simplify our analysis, we will focus on the regime of an asymptotically large dataset. In particular, we assume that trajectory pairs are sampled iid from a distribution d. As the dataset grows, the loss function (7) becomes

$$\mathcal{L}_\infty(\tilde{r}|d, r, V_*) = -\mathbb{E}_{(h,h')\sim d}\big[p_*(h, h'|r, V_*) \ln \tilde{p}(h, h'|\tilde{r}) + p_*(h', h|r, V_*) \ln \tilde{p}(h', h|\tilde{r})\big]. \quad (8)$$

We say a distribution d over trajectory pairs *compares transitions* if, for each $(h, h') \in \text{supp}(d)$, where $h = (s_0, a_0, \ldots, s_{T-1})$ and $h' = (s'_0, a'_0, \ldots, s'_{T'-1})$, $s_0 = s'_0$ and $T = T' = 2$. In other words, each trajectory pair elicits comparison between two different transitions from the same state. Consider a graph with vertices $\mathcal{S}$ and edges including all pairs (s_1, s'_1) such that $((s_0, a_0, s_1), (s_0, a'_0, s'_1)) \in \text{supp}(d)$ for some $s_0 \in \mathcal{S}$ and $a_0, a'_0 \in \mathcal{A}$. We say d *adjoins* s and s' if the two states are adjacent in this graph. We say d *connects* s and s' if the two states are connected in this graph.

The following result establishes conditions under which value is treated as reward (see appendix in extended version for the proof).

Theorem 2. (conflation from reward learning) *Consider an MDP $(\mathcal{S}, \mathcal{A}, P)$ and a reward function $r \in \Re^{\mathcal{S}}$. Let d be a distribution over trajectory pairs that compares transitions and connects all states. Let $\hat{r} \in \arg\min_{\tilde{r}\in\Re^{\mathcal{S}}} \mathcal{L}_\infty(\tilde{r}|d, r, V_*)$. Then, $\hat{r} - V_*$ is a constant function.*

That $\hat{r} - V_*$ is a constant function implies that preferences expressed by $\hat{r}$ are identical to those expressed by treating V_* as the reward function.

4.5 Misalignment

Suppose that the human attributes values to states based on V_*. The following result establishes that this gives rise to severe misalignment (see appendix in extended version for the proof).

Theorem 3. (severe misalignment from reward learning) *Consider the canonical example formulated in Section 3.2. Let $\hat{r} \in \arg\min_{\tilde{r}\in\Re^{\mathcal{S}}} \mathcal{L}_\infty(\tilde{r}|d, r, V_*)$ for a distribution d that compares transitions and connects all states. Then, for all $\epsilon \in (0, 1)$, $M > (1 + \epsilon + \epsilon^2)/(1 - \epsilon^2)$ and $\hat{\pi} \in \arg\max_\pi \hat{r}_\pi$, we have $r_{\hat{\pi}} = -1$.*

As trajectories are made longer, the standard reward learning algorithm will not recover the value function exactly. Instead, it will recover some confluence of the reward and value function. Severe misalignment ensues.

It is possible that *future* AI systems will be designed to disentangle what human choices convey about ends versus means. However, *perfect* disentanglement may be too difficult, and as we observed in our canonical example, even slight conflation can cause severe misalignment.

5 Manifestation in Real Environments

Our canonical example distills properties of the environment that make reinforcement learning highly sensitive to conflation of reward and value. While the example is simple, we expect the issue to arise across a broad range of more complex environments. Intuitively, the example highlights two properties that make reinforcement learning sensitive to conflation:

- P1. There are states that offer high reward but cannot be visited frequently.
- P2. There are states that offer high value and can be visited frequently but generate rewards well below average.

In our canonical example, the terminal goal offers high reward. The fact that it transitions to the common state, which imposes a long sojourn time, prevents frequent visits. The instrumental goal offers high value but reward well below average. That state can be visited frequently since self-transitions are allowed.

Theorem 1 applies only to our canonical example. We leave for future work developing a more general theorem that applies to a broad range of complex environments that exhibit the two properties listed above. In this section, we discuss a few examples of more complex and realistic environments to illustrate how these properties can make reinforcement learning sensitive to conflation more broadly. An important caveat is that our definition of conflation Definition 1 is likely to be too restrictive to hold in real settings. In practice, while the learned reward function is unlikely to be equivalent to a convex combination of the reward function and value function, we expect anticipation of future reward to heavily influence the learned reward function. In the examples in this section, there is a very clear intuitive sense in which reward and value is conflated. How best to formally characterize this sort of conflation is left for future work.

5.1 Arcade Games

In a study by Ibarz et al. (2018b), a reinforcement learning agent is trained on Atari games using learned reward functions. These functions tend to be highly shaped and assign proxy reward where true rewards are anticipated rather than immediately realized.

In one of the games, namely Montezuma's Revenge, the agent must obtain a key by traversing a ladder in order to complete the level. Once the level is completed, the game ends and no further reward can be earned. Thus, revisiting the goal state is not just difficult, but impossible.

In contrast, reaching the top rung of the ladder, which represents an instrumental goal, is easy and can be repeated any number of times. Because the learned reward function attributes proxy reward to high-value states, the trained agents remains at the top rung indefinitely (see Figure 3).

A similar phenomenon occurs in the game Private Eye. Here, the agent repeatedly moves left and right below a window where a suspect is hiding. Instead of jumping up to the window to catch the suspect, the agent continues to move left and right, accruing reward according to the learned reward model but not the game. Sometimes the agent does jump up towards the suspect but misses the suspect. This suggests that the learned reward function assigns proxy reward to 'almost catching the suspect'. This is a separate phenomenon from reward-value conflation.

While these outcomes are not catastrophic — especially since they are in the context of video games — they illustrate a dynamic that could have more serious consequences in real environments.

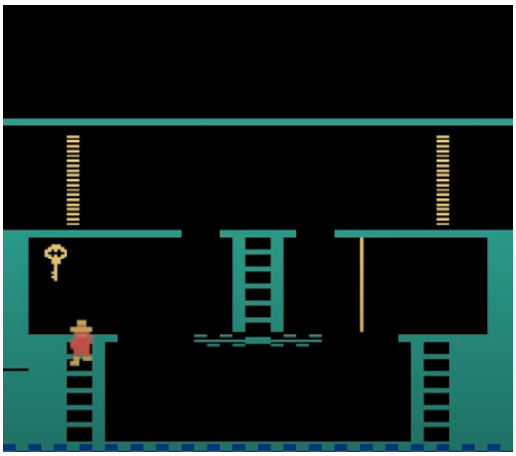


Figure 3: Stuck at the top rung of a ladder in Montezuma's revenge. From a video linked in Ibarz et al. (2018b), see Ibarz et al. (2018a). The video shows the proxy reward over time: it is high when the agent moves up the ladder, suggesting reward value conflation. Further, the misaligned agent remains on the ladder instead of progressing through the game.

5.2 AI Therapist

The following hypothetical example serves to illustrate how a similar issue can arise in a practical application. Consider an AI therapist trained to treat patients with obsessive compulsive disorder (OCD). These patients engage in unproductive behaviors such as obsessive handwashing or checking of locks. The goal is to help patients overcome these tendencies.

In human-delivered therapy, OCD treatment often takes the form of exposure therapy (Hezel and Simpson 2019): the patient is exposed to thoughts that trigger their obsessive behaviors and then asked to abstain from their compulsion. Initially, the patient may be asked to abstain from the behavior for a short duration — perhaps a minute — before engaging in it. Over time, the intensity of the exposure therapy is increased, for instance, by asking the patient to abstain for longer durations.

Suppose a proxy reward function is learned from choices between professional transcripts, each generated by a professional therapist. Due to aforementioned flaws in common approaches to reward learning, the resulting proxy is likely to reward instrumental goals such as having the patient abstain over a short duration. Further, if the AI therapist successfully cures the patient, the patient no longer needs the AI therapist. Therefore, once the patient is cured, the AI therapist cannot accrue further proxy rewards. These observations give rise to the dynamic of concern: a policy that maximizes proxy reward might repeatedly induce short abstentions and never build up to longer durations as required to cure the patient. In particular, with such a policy, the AI therapist continually accrues proxy reward.

Obviously, upon observing this behavior, a well-intended provider of an AI therapy product would fix this issue with a bespoke solution. But then there might be yet another instrumental goal pursued by the AI therapist that calls for yet another bespoke solution. It would be better to have a principled and robust approach that does not rely on addressing particular bad behaviors as they are observed.

5.3 Shutdown Evasion

A long-standing concern is that AI systems may resist shutdown (Clarke 1968; Soares et al. 2015; Hadfield-Menell et al. 2017; Wängberg et al. 2017; Turner et al. 2019; Nolan 2025; Rosenblatt 2025). The argument for concern goes as follows. Consider an AI system that is well described as trying to achieve some goal. For example, it may be a reinforcement learning agent that is optimizing a reward function. Then, for a wide range of reward functions, staying alive is instrumentally useful since most goals are impossible to achieve if the agent has shut down (Bostrom 2012; Omohundro 2018). As Russell (2022) says, "you can't fetch the coffee if you are dead". In this case, the worry is not that the agent treats self-preservation as a terminal goal. Rather, the worry is that self-preservation is a means to a wide range of ends.

Here, we highlight a different mechanism in which the agent treats self-preservation as an end in itself. Suppose the human's terminal goal is for the agent to shut down. Then, it will be instrumentally useful to take steps *towards* being shut down. Then, if the agent conflates reward and value it will accrue some proxy reward just for taking steps towards being shut down. But then, perversely, the agent may be incentivized to stay on, so that it can accrue such proxy rewards indefinitely.

6 Related Work

There is a substantial literature discussing how reward functions that encode instrumental goals can incentivize suboptimal behavior (see e.g., (Randløv and Alstrøm 1998; Sutton and Barto 1998; Ho et al. 2015; Amodei and Clark 2016; Ibarz et al. 2018b). Sutton and Barto (1998, Chapter 3, p. 58) say that *"the reward signal is not the place to impart to the*

agent prior knowledge about how to achieve what we want it to do... If achieving ... subgoals were rewarded, then the agent might find a way to achieve them without achieving the real goal." Russell and Norvig (2016, Chapter 2) provide similar advice.

There are many recorded examples where manually specified reward functions encode not only terminal goals but also instrumental goals, leading to unintended consequences. A well-known case is the boat-race environment, in which the intention is for the agent to complete laps as quickly as possible (Amodei and Clark 2016). To incentivize this, the agent receives reward for hitting intermediate targets. As an unintended consequence, the agent learns to drive in circles, repeatedly hitting the same targets, without completing the lap. In the same spirit, Randløv and Alstrøm (1998) study an agent that steers a bicycle that is rewarded for moving closer to the goal rather than for reaching it. The agent learns to circle the goal at varying distances, accruing reward indefinitely without ever completing the task.

Although attributing reward to instrumental goals can give rise to misalignment, it can also accelerate learning by providing denser reward signals (Singh, Lewis, and Barto 2009; Singh et al. 2010). In this context, specifying a reward function that rewards progress is often known as *reward shaping*. Because reward shaping can speed up learning, substantial effort has been dedicated to develop methods that do that without inducing negative outcomes (Ng, Harada, and Russell 1999; Randløv and Alstrøm 1998; Wiewiora, Cottrell, and Elkan 2003; Asmuth, Littman, and Zinkov 2008; Knox and Stone 2009; Devlin and Kudenko 2012; Grześ 2017; Zou et al. 2019; Devidze, Kamalaruban, and Singla 2022; Lidayan, Dennis, and Russell 2024). The classical paper by Ng, Harada, and Russell (1999) provides a way of constructing shaped reward functions that maintain preferences among policies.

In the context of RLHF, multiple papers discuss how human feedback often reflects not only terminal goals but also how to achieve those goals (Thomaz, Breazeal et al. 2006; Knox et al. 2012; Ho et al. 2015, 2019; Ibarz et al. 2018b; Knox et al. 2022, 2024; Marklund and Van Roy 2024). Due to a mismatch between the assumed and the actual human feedback model, reward learning can result in reward functions that encode how to achieve goals rather than the terminal goals themselves (see, e.g., Thomaz, Breazeal et al. (2006); Knox et al. (2012); Ho et al. (2015); Gong and Zhang (2020)).

Ho et al. (2015) gives an example. In their study, they find that humans often give feedback indicative of action quality (how good the action is relative to the optimal action) rather than immediate reward. In that case, when feedback is interpreted as immediate reward, they show that it can give rise to a misaligned reward function that incentivizes an agent to remain within unrewarding states. This insight anticipates how now common approaches to reward learning treat means as ends, as well as how in our canonical example misspecification (see Section 3.2) leads an agent to loop endlessly at the instrumental goal.

In Knox et al. (2024), they also study learned reward functions that assign proxy reward based on action-quality rather than true reward. In particular, they study what happens when the learned reward function is the *optimal advantage function* or an approximation thereof. The optimal advantage of a state s and an action a is defined by $A_*(s,a) = Q_*(s,a) - V_*(s)$ where Q_* and V_* are the optimal action-value and value function, respectively. We showed that using the value function as the proxy reward function (i.e. maximal conflation) can lead to severe misalignment. In contrast, Knox et al. (2024) note that optimal policies are preserved when the optimal advantage function is used as the proxy reward function. Still, they find empirically that when learned reward only approximates optimal advantage, a reinforcement learning agent maximizing learned reward fares poorly in a set of episodic environments.

There is also work on evolved reward functions (see, e.g., Singh, Lewis, and Barto (2009); Singh et al. (2010)), which explains that these functions often encode not only terminal goals but also instrumental goals, as this speeds up learning. In evolutionary biology, this is often used to explain why desires for food and play seem to be hardwired even though they are only indirectly related to genetic fitness.

The results we have presented in this paper extend these prior works by identifying specific conditions on the environment and the human's preferences under which even slight conflation of instrumental and terminal goals gives rise to misalignment. While previous work also emphasizes the importance of having the reward function encode only terminal goals, our work identifies a specific mechanism that gives rise to misalignment even when conflation of instrumental and terminal goals is slight.

7 Closing Remarks

We established that conflating instrumental and terminal goals, even slightly, can give rise to severe misalignment. Two properties render an environment susceptible to this failure mode. The first is that states with high reward cannot be visited frequently. The second is the existence of states that offer high value and can be visited frequently but generate rewards well below average. These observations motivate future work to investigate how this failure mode manifests empirically in real environments and how it can be mitigated.

Acknowledgements

We thank Alex Cloud, Alex Turner, Anmol Kagrecha, Jonathan Carr, Saurabh Kumar and Stephane Hatgis-Kessell for helpful discussions and feedback. Alex Infanger thanks Bryce Woodworth for his ML Alignment & Theory Scholars (MATS) research management support. This research was in part supported by a grant from the US Army Research Office. Alex Infanger was financially supported by the MATS program.

References

Amodei, D.; and Clark, J. 2016. Faulty reward functions in the wild. OpenAI Blog. https://openai.com/blog/faulty-reward-functions/. Accessed: 2025-12-06.

Asmuth, J.; Littman, M. L.; and Zinkov, R. 2008. Potential-based Shaping in Model-based Reinforcement Learning. In *AAAI*, 604–609.

Blackwell, D. 1962. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 719–726.

Bostrom, N. 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22: 71–85.

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

Clarke, A. C. 1968. *2001: A Space Odyssey*. New American Library.

Devidze, R.; Kamalaruban, P.; and Singla, A. 2022. Exploration-guided reward shaping for reinforcement learning under sparse rewards. *Advances in Neural Information Processing Systems*, 35: 5829–5842.

Devlin, S. M.; and Kudenko, D. 2012. Dynamic potential-based reward shaping. In *11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, 433–440. IFAAMAS.

Gong, Z.; and Zhang, Y. 2020. What is it you really want of me? generalized reward learning with biased beliefs about domain dynamics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2485–2492.

Grześ, M. 2017. Reward Shaping in Episodic Reinforcement Learning. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, 565–573. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Hadfield-Menell, D.; Dragan, A. D.; Abbeel, P.; and Russell, S. 2017. The Off-Switch Game. In *AAAI Workshops*.

Hezel, D. M.; and Simpson, H. B. 2019. Exposure and response prevention for obsessive-compulsive disorder: A review and new directions. *Indian journal of psychiatry*, 61(Suppl 1): S85–S92.

Ho, M. K.; Cushman, F.; Littman, M. L.; and Austerweil, J. L. 2019. People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, 148(3): 520.

Ho, M. K.; Littman, M. L.; Cushman, F.; and Austerweil, J. L. 2015. Teaching with rewards and punishments: Reinforcement or communication? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 37.

Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018a. Hacking a trained reward model: Montezuma's Revenge. https://www.youtube.com/watch?v=_sFp1ffKIc8&list=PLehfUY5AEKX-g-QNM7FsxRHgiTOCl-1hv&index=2. Accessed: 07-31-2025.

Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018b. Reward learning from human preferences and demonstrations in Atari. *Advances in neural information processing systems*, 31.

Knox, W. B.; Glass, B. D.; Love, B. C.; Maddox, W. T.; and Stone, P. 2012. How humans teach agents: A new experimental perspective. *International Journal of Social Robotics*, 4: 409–421.

Knox, W. B.; Hatgis-Kessell, S.; Adalgeirsson, S. O.; Booth, S.; Dragan, A.; Stone, P.; and Niekum, S. 2024. Learning optimal advantage from preferences and mistaking it for reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10066–10073.

Knox, W. B.; Hatgis-Kessell, S.; Booth, S.; Niekum, S.; Stone, P.; and Allievi, A. 2022. Models of human preference for learning reward functions. *arXiv preprint arXiv:2206.02231*.

Knox, W. B.; and Stone, P. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*, 9–16.

Lidayan, A.; Dennis, M.; and Russell, S. 2024. BAMDP shaping: a unified theoretical framework for intrinsic motivation and reward shaping. *arXiv preprint arXiv:2409.05358*.

Marklund, H.; and Van Roy, B. 2024. Choice Between Partial Trajectories: Disentangling Goals from Beliefs. arXiv:2410.22690.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, 278–287. Citeseer.

Nolan, B. 2025. Anthropic's new AI model threatened to reveal engineer's affair to avoid being shut down. *Fortune*.

Omohundro, S. M. 2018. The basic AI drives. In *Artificial intelligence safety and security*, 47–55. Chapman and Hall/CRC.

Randløv, J.; and Alstrøm, P. 1998. Learning to Drive a Bicycle Using Reinforcement Learning and Shaping. In *ICML*, volume 98, 463–471. Citeseer.

Rosenblatt, J. 2025. AI Is Learning to Escape Human Control. *The Wall Street Journal*.

Russell, S. 2022. Human-Compatible Artificial Intelligence.

Russell, S. J.; and Norvig, P. 2016. *Artificial intelligence: a modern approach*. pearson.

Singh, S.; Lewis, R. L.; and Barto, A. G. 2009. Where do rewards come from. In *Proceedings of the annual conference of the cognitive science society*, 2601–2606. Cognitive Science Society.

Singh, S.; Lewis, R. L.; Barto, A. G.; and Sorg, J. 2010. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2): 70–82.

Soares, N.; Fallenstein, B.; Armstrong, S.; and Yudkowsky, E. 2015. Corrigibility. In *AAAI Workshop: AI and Ethics*.

Sutton, R. S.; and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Thomaz, A. L.; Breazeal, C.; et al. 2006. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Aaai*, volume 6, 1000–1005. Boston, MA.

Turner, A. M.; Smith, L.; Shah, R.; Critch, A.; and Tadepalli, P. 2019. Optimal policies tend to seek power. *arXiv preprint arXiv:1912.01683*.

Wängberg, T.; Böörs, M.; Catt, E.; Everitt, T.; and Hutter, M. 2017. A Game-Theoretic Analysis of the Off-Switch Game. In Everitt, T.; Goertzel, B.; and Potapov, A., eds., *Artificial General Intelligence*, 167–177. Cham: Springer International Publishing.

Wiewiora, E.; Cottrell, G. W.; and Elkan, C. 2003. Principled methods for advising reinforcement learning agents. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, 792–799.

Zou, H.; Ren, T.; Yan, D.; Su, H.; and Zhu, J. 2019. Reward shaping via meta-learning. *arXiv preprint arXiv:1901.09330*.