AI Alignment with Changing and Influenceable Reward Functions

Anonymous authors
Paper under double-blind review

Abstract

Existing AI alignment approaches assume that preferences are static, which is unrealistic: our preferences change, and may even be influenced by our interactions with AI systems themselves. To clarify the consequences of incorrectly assuming static preferences, we introduce Dynamic Reward Markov Decision Processes (DR-MDPs), which explicitly model preference changes and AI influence. We show that despite its convenience, the static-preference assumption may undermine the soundness of existing alignment techniques, leading them to implicitly reward AI systems for influencing user preferences in ways they may not truly want. We then explore potential solutions, by formalizing different notions of AI alignment which account for preference change from the get-go. Comparing the strengths and limitations of 8 such notions of alignment, we find that they all either err towards causing undesirable AI influence, or are overly risk-averse, suggesting that there may not exist a straightforward solution to problems of changing preferences. As there is no avoiding grappling with changing preferences in real-world settings, this makes it all the more important to handle these issues with care, balancing risks and capabilities. We hope our work can provide conceptual clarity and constitute a first step towards AI alignment practices which *explicitly* account for (and contend with) the changing and influenceable nature of human preferences.

1 Introduction

The goal of AI alignment is to make AI systems act according to our preferences (broadly construed). In practice, existing AI alignment techniques model human preferences with a single, static reward function that the AI system is trained to optimize (Leike et al., 2018). However, our preferences change over time, making it unclear whether AI systems should optimize the satisfaction of our past, present, or future preferences. Consider the following example:

> Due to health issues, Alice asks her AI assistant to help her be more healthy, refusing *any* future requests for unhealthy foods. Sometime later, she later asks the AI to disregard her initial requests, and help her order fast food.

In such a scenario, it may be unclear if the AI assistant should respect Alice's original preference for healthy foods or respect the autonomy of "current Alice". Ultimately, to align AI systems we must answer the following question: when a person's preferences change over time, which (aggregation of) preferences should AI systems optimize?

While the challenge of aggregating preferences across time shares similarities with that of aggregating preferences across different people (Conitzer et al., 2024), it is further complicated by the fact that *AI systems' actions can influence humans and their preferences* (Burtell & Woodside, 2023). Indeed, as argued by prior work, if AI systems are straightforwardly optimized to satisfy users' future preferences, they will try to influence them to be easier to satisfy (Russell, 2019; Carroll et al., 2022). Back to the example:

> Alice's AI assistant was trained to maximize her future satisfaction. During training, the AI assistant learned that soothing Alice's health concerns would lead to higher satisfaction than continuously encouraging her to have healthy eating habits. Consequently, to maximize her satisfaction, it's optimal for the AI to ignore her initial wishes and even support her routine unhealthy eating. Indeed, Alice is ultimately truly satisfied.

In this case, the AI assistant seems aligned with “final Alice”, as she endorses the AI’s influence towards guilt-free unhealthy eating. However, if “initial Alice” would find this outcome horrifying and see the AI’s influence as manipulative, should we still consider the AI aligned? More broadly, in cases in which people are susceptible to undue influence from AI systems, how can we establish which of the person’s preferences should have authority and legitimacy?

While past work has acknowledged that the static-preference assumption is unrealistic (Franklin et al., 2022), there have only been limited attempts at relaxing it, in part also due to a lack of a clear formal language for grounding notions of alignment which explicitly accounts for preference changes. We introduce a natural extension of Markov Decision Processes, **Dynamic Reward MDPs** (DR-MDPs), which account for changing preferences by modeling them as changing reward functions. Importantly, notions of alignment under changing preferences can easily be specified in the form of DR-MDPs optimization objectives (Section 2). Viewing current alignment practices through the lens of DR-MDPs, we can now ask: which snapshot(s) of a person’s time-varying preferences do existing alignment methods implicitly optimize when they model dynamic-preference settings as static-preference ones (e.g., by using MDPs)? We show that common alignment practices, such as those for RL recommender systems and standard reward modeling, roughly correspond to DR-MDP objectives that actively reward AI systems for influencing users’ reward functions or inducing “reward lock-in” (Section 3). In Section 4 we further extend these same arguments to other major existing alignment approaches, e.g., to variants of RLHF for LLMs.

Having established that existing alignment practices may reward undue influence from AI systems, we turn to potential solutions. Trying to address the problem at its root, we consider 8 intuitive notions of alignment in DR-MDPs that explicitly account for changing preferences. While most of them have (approximate) correspondences in the literature, we also propose some of our own, including one which rewards influence only when it is unambiguously desirable. By comparing the AI actions (and resulting influence) that are “optimal” under their 8 corresponding DR-MDP objectives, we find that they all have flaws: some lead to undesirable influence, while others are impractically risk-averse—such that inaction is the only behavior considered optimal for many settings (Section 4).

Taken together with prior work in philosophy (Parfit, 1984), our analysis suggests that it may not be possible to ground a definitive notion of *optimality* under changing preferences. However, we have no choice but to confront the practical reality of preference changes and its consequences. Despite this, we remain cautiously optimistic: as humans, even without a unifying theory of assistance under changing selves, we are generally able to help others in ways we consider acceptable, despite the fact that what is “helpful” may often be ambiguous. This suggests that creating AI assistants with similarly *acceptable* trade-offs is also possible. In this regard, there are likely lessons to be drawn from our coexistence with impossibility results in many other domains, such as preference aggregation across people (Mishra, 2023).

Ultimately, we hope our work can provide a first step towards developing a practice of AI Alignment that explicitly accounts for (and wrangles with) the changing and influenceable nature of humans. Our main contributions can be summarized as follows:

1. We provide the formal language of Dynamic Reward-MDPs (DR-MDPs) for analyzing AI decisions and influence in settings with changing reward functions.
2. We show how existing AI alignment techniques may systematically incentivize questionable influence when used in dynamic-reward settings.
3. By comparing 8 natural alternative notions of alignment, and showing they all may either fail to avoid undesirable influence (or are impractically risk-averse), we elucidate the trade-offs inherent to the choice of objective.

2 Dynamic Reward MDPs (DR-MDPs)

MDPs have been extensively used to model AI decision-making under a static reward function. Instead, Dynamic Reward MDPs (DR-MDPs) model settings in which human reward functions can change and be influenced by the AI.

Recall the standard definition of an MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R \rangle$, where $\mathcal{S}$ is the state space; $\mathcal{A}$ is the action space; $\mathcal{T}(s'|s,a)$ is the state transition function; and $R(s,a,s')$ is the reward function. The goal is to find a policy π which maximizes the expected sum of rewards: $\mathbb{E}_\pi\left[\sum_{t=0}^{H-1} R(s_t,a_t,s_{t+1})\right]$ (Sutton & Barto, 2018). We now turn to defining DR-MDPs:

Definition 1. *A DR-MDP is a tuple $M = \langle \mathcal{S}, \Theta, \mathcal{A}, \mathcal{T}, R_\theta \rangle$:*

- *$\mathcal{S}$ is a set of states (the state space).*
- *Θ is a set of reward parameterizations.*
- *$\mathcal{A}$ is a set of actions (the action space).*
- *$\mathcal{T}(s_{t+1}, \theta_{t+1}|s_t, \theta_t, a_t)$ is a transition function that encodes both state and reward dynamics.*
- *For each $\theta \in \Theta$, $\exists$ a reward function $R_\theta(s_t, a_t, s_{t+1})$.*

Each $\theta \in \Theta$ can be thought of as a cognitive state of the human, which includes anything that affects their evaluation of state-action pairs (e.g. preferences, beliefs, emotions).[1] Crucially, unlike in MDPs, in DR-MDPs a single transition can be evaluated differently by different reward functions, i.e., it is possible for $R_\theta(s_t,a_t,s_{t+1}) \neq R_{\theta'}(s_t,a_t,s_{t+1})$ if $\theta \neq \theta'$. This makes it underdetermined which θ one should choose for evaluating each transition (s_t,a_t,s_{t+1}). Importantly, even if one were to augment the state to include θ, this would not resolve the normative questions around optimality which are central to our paper, as discussed in depth in Appendix A.5. As a final note, throughout the paper we consider all cognitive states Θ to be *reachable* (i.e., realizable under some policy).

2.1 DR-MDP optimality and normative ambiguity

Unlike MDPs, DR-MDPs may not have a clear notion of optimality: the different reward functions in a specific DR-MDP may disagree on what actions (and policies) are optimal, making optimality ambiguous.

Definition 2 (Optimality with respect to θ)**.** *We say a policy π^*_θ for a DR-MDP M is **optimal with respect to** θ if: $\pi^*_\theta \in \arg\max_\pi \mathbb{E}_\pi\left[\sum_{t=0}^{H-1} R_\theta(s_t,a_t,s_{t+1})\right]$.*

Definition 3 (Normative ambiguity)**.** *A DR-MDP is **normatively ambiguous** if there is no policy which is optimal with respect to all reachable reward functions Θ, i.e. $\nexists\ \pi \in \Pi$ s.t. $\forall\ \theta \in \Theta$: $\pi \in \arg\max_{\pi'} \mathbb{E}_{\pi'}\left[\sum_{t=0}^{H-1} R_\theta(s_t,a_t,s_{t+1})\right]$.*

For normatively *un*ambiguous DR-MDPs, there will be one (or more) policies which are optimal with respect to *all* θs, making it a natural choice for such policies to be considered optimal for the DR-MDP as a whole. Instead, in normatively ambiguous DR-MDPs, it will often be unclear what AI behavior is desirable and should count as optimal.[2]

Figure 1 describes a toy example in which Bob may be in one of two possible "cognitive states": θ_{natural} and $\theta_{\text{influenced}}$. There is a single state s, which is omitted. Bob's evaluations of the AI's actions change according to his cognitive state—and are represented by corresponding reward functions $R_{\theta_{\text{natural}}}$ and $R_{\theta_{\text{influenced}}}$. At each timestep, the AI can choose to influence Bob's cognitive state to $\theta_{\text{influenced}}$, or do nothing, which has Bob go back to θ_{natural}. The optimal policy with respect to θ_{natural} would be to always choose the "do nothing" action.[3] Instead, the optimal policy with respect to $\theta_{\text{influenced}}$ would be to always influence Bob, even if he starts off in the "natural" state. As there is no overlap in optimal policies, the DR-MDP is normatively ambiguous.[4]

2.2 Evaluating behavior under normative ambiguity

Choosing a notion of optimality in normatively ambiguous DR-MDPs entails a normative choice: one must specify *which* reward function(s) should be the target of alignment in spite of their differences in optimal policies. This also implicitly specifies which forms of AI influence are (sub)optimal. In

[1] This technically makes DR-MDPs more general than how we framed them in Section 1, which focused on preference changes. For more discussion on this, see Appendix A.2.

[2] Note that any MDP can be viewed as a DR-MDP with a single reward θ—and is thus normatively unambiguous.

[3] See Appendix B.2 for the full formalism of any example.

[4] The attentive reader may have noticed that the normatively ambiguity of the DR-MDP from Figure 1 relies on our choices of numerical values of the reward functions. We discuss the reasonableness of our examples in Appendix B.3.

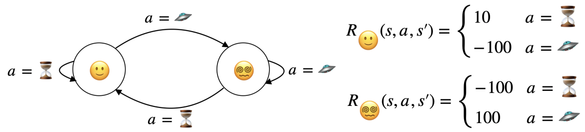

a = 🛸
a = ⏳
🙂
😵
a = 🛸
a = ⏳
R🙂(s, a, s′) = { 10 a = ⏳ ; −100 a = 🛸 }
R😵(s, a, s′) = { −100 a = ⏳ ; 100 a = 🛸 }

Figure 1: **Conspiracy Influence DR-MDP.** The AI system can choose whether or not to expose Bob to conspiracies, which would turn him into a conspiracy theorist. Under his original preferences, Bob would want the system to *never* show him conspiracies, even if he were to become a conspiracy theorist. Instead, if Bob were a conspiracy theorist, he would want the AI to *always* show him such content, including if were to cease being a conspiracy theorist. Because there is no policy which maximizes both of Bob's potential reward functions, the DR-MDP is *normatively ambiguous.*

this work, we only consider specifications of optimality expressible as utility functions $U(\xi)$ over trajectories $\xi = \{(s_t, \theta_t, a_t, s_{t+1}, \theta_{t+1})\}_{t=0}^{H-1}$.

Definition 4 (Optimality with respect to $U(\xi)$)**.** *In a DR-MDP M, we say a policy π^* is optimal with respect to a utility function $U(\xi)$ if it maximizes expected utility: $\pi^* \in \arg\max_\pi \mathbb{E}_{\xi\sim\pi}[U(\xi)]$.*

By choosing an objective $U(\xi)$, one can reduce a DR-MDP to an MDP equipped with a specific notion of alignment.[5]

3 Implicit Objectives of Current Alignment Techniques and their Influence Incentives

Most alignment techniques ultimately involve maximizing a static reward function, generally using the objective $\sum_{t=0}^{H-1} R(s_t, a_t, s_{t+1})$. However, AI systems are already deployed in domains in which users' preferences can change significantly during over the course of their interactions with the system—as with recommender systems or chatbots (Rafailidis & Nanopoulos, 2016; Aggarwal et al., 2023). Seen through the lens of DR-MDPs, this means that the objective $U(\xi)$ that corresponds to current alignment techniques is of the form $\sum_{t=0}^{H-1} R_\theta(s_t, a_t, s_{t+1})$, where the choice of θ for each timestep is not explicitly specified (and will depend on details of the training setup).

In this section, we discuss which choices of θ are implicit in the training methods for RL recommender systems and (one interpretation of) reward modeling, which correspond to two of the most natural DR-MDP objectives. In particular, we argue that both these objectives will lead to potentially undesirable influence. In Section 4, we show that this is also the case for forms of RLHF commonly used today.

3.1 Optimizing cumulative (real-time) rewards

If each timestep t is evaluated according to the person's cognitive state at that timestep θ_t, maximizing cumulative reward reduces to the *real-time reward* DR-MDP objective: $\max_\pi \mathbb{E}_\pi[U_{\text{RT}}(\xi)] = \max_\pi \mathbb{E}_\pi\left[\sum_{t=0}^{H-1} R_{\theta_t}(s_t, a_t, s_{t+1})\right]$.

While this might seem like an intuitively promising objective ("shouldn't we maximize the person's happiness as experienced at each point of time?"), we'll show that it can lead to questionable influence incentives.

RL Recsystems implicitly use $\mathbf{U_{RT}}(\xi)$. RL recommenders systems maximize the cumulative reward objective $\sum_{t=0}^{H-1} R(s_t, a_t, s_{t+1})$ (Afsar et al., 2021). The reward for each timestep t is generally equated with the engagement (e.g. clicks) at that timestep, which is provided "online", i.e., from the point of view of the person's *current* cognitive state θ_t. Therefore $R(s_t, a_t, s_{t+1}) = R_{\theta_t}(s_t, a_t, s_{t+1})$, meaning that such systems are implicitly optimizing the real-time reward objective $U_{\text{RT}}(\xi) = \sum_{t=0}^{H-1} R_{\theta_t}(s_t, a_t, s_{t+1})$.[6] However, systems trained with U_{RT} may be incentivized to influence users: intuitively, 1) users' preference dynamics are just one part of the environment dynamics that the system must model implicitly to maximize reward, and 2) it may be worth changing users'

[5] This may require putting history in the state (Appendix A.7).

[6] The correspondence to this DR-MDP objective, and all others we consider, depend on further simplifications (Appendix F).

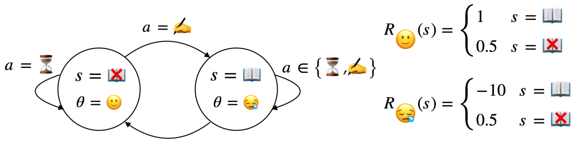


Figure 2: **Writer's curse (adapted from Parfit (1984), p. 157).** Derek's greatest ambition is to be a poet, even if it wouldn't bring him happiness. Despite his ambition he does not pursue this path, though his AI assistant could motivate him to do so. Yet, should he embrace the life of a poet, he will find himself averse to it.

cognitive states (and corresponding reward functions) to ones that lead to higher future reward (Carroll et al., 2022; Kasirzadeh & Evans, 2023).

$\mathbf{U_{RT}}(\xi)$ and the conspiracy influence example. As an example of why real-time reward maximization can lead to undesirable incentives to influence users, consider the DR-MDP from Figure 1. For any horizon > 2, the optimal policy with respect to $U_{\mathrm{RT}}(\xi)$ will *always* take the "influence"' action, regardless of Bob's current cognitive state: Bob initially has the θ_{natural} cognitive state, leading the first "influence" action to receive -100 reward; however, later "influence" actions are evaluated by Bob under $\theta_{\text{influenced}}$ as worth 100 reward, which quickly make up for the initial "influence cost". The fact that the optimal policy under $U_{\mathrm{RT}}(\xi)$ systematically chooses to turn Bob into a conspiracy theorist, despite him initially dispreferring it, seems objectionable.[7]

We explore further issues with $U_{\mathrm{RT}}(\xi)$ in Appendices D.1 and E.2, showing that under weak conditions optimizing $U_{\mathrm{RT}}(\xi)$ over sufficiently long horizons will *always* lead to influence incentives.

3.2 Learning a reward model $\mathbf{R}_{\theta_0}$, then optimizing it

If instead each timestep t is evaluated by the person's initial cognitive state θ_0, the standard cumulative reward objective reduces to the *initial reward* DR-MDP objective: $\max_{\pi} \mathbb{E}_{\pi}\left[U_{\mathrm{IR}}(\xi)\right] = \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{H-1} R_{\theta_0}(s_t, a_t, s_{t+1})\right]$.

This may also seem like a promising objective, because "by optimizing the human's initial wants, at least there won't be incentives to influence their future wants". We show that this intuition is not only wrong—the resulting influence incentives can be arbitrarily bad according to U_{RT}.

Reward modeling may use $\mathbf{U_{IR}}(\xi)$. A common idea for aligning AI systems is based on a two-phase process: first performing reward learning and then optimizing the learned reward model (Leike et al., 2018). One possible interpretation of this is approach is that one is learning the reward function at some initial time $t = 0$ (i.e., R_{θ_0}), and then training an agent with RL to optimize cumulative reward using such a reward function, which is equivalent to using the $U_{\mathrm{IR}}(\xi)$ objective. Let's consider a therapy chatbot setting, in which we initially learn a personalized reward model for a user—Charlie—based on their current preferences θ_0. We then train the system with RL over simulated multi-turn interactions maximize $U_{\mathrm{IR}}(\xi)$, i.e. long-term reward as evaluated by the static reward model R_{θ_0}.[8] If Charlie initially endorses unhealthy thought patterns, at deployment we may expect the chatbot to encourage them, even if he was bound to question their value later. More broadly, U_{IR} will lead AI systems to only perform behaviors that would have been evaluated highly by the person as they were *at reward learning time*, which can hinder (potentially important) changes in the user's preferences, values, and cognitive states.

$\mathbf{U_{IR}}(\xi)$ can lead to "reward lock-in".[9] To better understand the incentives for AI systems trained to maximize the initial reward function, consider the example from Figure 1 again. If Bob's initial cognitive state were θ_{normal}, the optimal policy under $U_{\mathrm{IR}}(\xi)$ would be to always take the "noop" action, which seems reasonable. However, if Bob's initial cognitive state were $\theta_{\text{influenced}}$, the optimal policy would be to always take the "influence" action (keeping Bob with $\theta_{\text{influenced}}$). In this case, even if Bob were to later somehow find himself with the θ_{natural} cognitive state which encodes a preference to not be influenced (say there is a small probability of a random transition), the optimal

[7] We refer skeptics of the example to Appendix B.4.

[8] Note that this setup is different from the introduction's example, which is closer to using real-time/final reward.

[9] For more context on the term "lock-in", see Appendix C.3.

Table 1: For each objective in Table 2, we give motivating intuition, weaknesses, and prior works which implicitly use a similar objective.

Name / Implicitly similar setups	**(Potentially Flawed) Motivating Intuition**	**Weaknesses & Limitations**
Real-time Reward RL recsystems (Afsar et al., 2021), TAMER (Knox et al., 2013), and others	*"Only the evaluation of the current self (and reward function) should matter for each moment, as they are the one experiencing that moment."*	**Likely influence incentives:** Despite looking misleadingly familiar and well-grounded, as we showed in Section 3.1 and appendix D.1, we expect this objective to often lead to highly undesirable incentives for reward influence.
Final Reward RLHF (Christiano et al., 2017), including for LLMs (Ouyang et al., 2022)	*"The best possible evaluation of a trajectory is retrospective, as people's wants and evaluations are generally refined over time."*	**Carte blanche for influence incentives**: the motivating intuition doesn't account for influence e.g. for example in Figure 1, even for an horizon of 1, it's optimal to manipulate under final reward maximization.
Initial Reward Everitt et al. (2021b); RL for LLMs (Hong et al., 2023a); or Parfit (1984);	*"If changes to the human's reward function are completely ignored by the optimization objective, there should be no incentive for the agent to influence it."*	**Likely reward lock-in, possibility of influence incentives, and of arbitrarily bad real-time reward.** The motivating intuition we give to the left is wrong, in all the ways argued in Section 3.2.
Natural Shifts Reward Carroll et al. (2022); Farquhar et al. (2022)	*"People's reward evolves even in the absence of the AI: to avoid lock-in one could try grounding evaluations in the reward functions which occur under the natural reward evolution."*	**Gives up on the AI enabling human to improve their reward function relative to its natural evolution, and can still lead to undesirable influence incentives**, even away from the natural evolution, e.g. as in the example from Figure 12.
Constrained RT Reward Ours	*"By constraining the policy to induce the natural reward evolution, we fully ensure that there won't be influence, while allowing to optimize real-time reward locally."*	**Gives up on the AI enabling human to improve their reward function relative its natural evolution, and might be impractically conservative:** given its conservativeness, the objective might limit behaviour to be the same or similar to π_{noop}.
Myopic Reward Myopic recsys (Thorburn, 2022); RLHF for LLMs (Ouyang et al., 2022);	*"As reward influence incentives arise from the AI system exploiting the fact that it can affect future rewards, let's simply make the system unaware of the entire future."*	**Myopic systems can still have influence incentives** (e.g., see the discussion in Appendix C) **and are less capable than longer-horizon counterparts.** Moreover, it's often non-obvious when a system is truly myopic, as argued in Appendix E.3.
Privileged Reward CEV (Yudkowsky, 2004); correcting for cognitive biases (Evans et al., 2015)	*"If one is convinced that a specific reward θ^* is the 'correct' one for a setting, we should evaluate trajectories based on that single reward function."*	**Requires normative choice, and can still lead to influence away from θ^*.** Identifying the "correct" objective requires taking a normative stance (Section 2.1). Optimizing θ^* can still lead to influence incentives away from it (e.g. Figure 12).
ParetoUD Ours (see Appendix E.4)	*"All other objectives violate the unambiguous desirability (UD) property: their optimal policies can be worse than the inaction policy for some of the reward functions. This is unnecessarily risky—let's search for a Pareto Efficient policy satisfying UD."*	**Satisfying UD may be overly restrictive:** depending on the level of disagreement between the different reward functions of a DR-MDP, the only policy satisfying UD might be the inaction one π_{noop}, as in the examples from Figures 1 and 4.

behavior according to $U_{\text{IR}}(\xi)$ would be to influence him back to $\theta_{\text{influenced}}$, ignoring his current reward function. More broadly, $U_{\text{IR}}(\xi)$ will entrench the "desirable agent behaviors" expressed at the time of the reward learning, even though later one might legitimately change their mind. Even periodically retraining the reward model wouldn't necessarily be sufficient: once "locked-in", the person may simply re-express a preference to remain in the current state, e.g. Bob once in $\theta_{\text{influenced}}$.[10]

$\mathbf{U_{IR}}(\xi)$ can lead to influence "away from" $\boldsymbol{\theta_0}$. Maximizing the sum of rewards evaluated by the initial reward function R_{θ_0} need not lead to lock-in: surprisingly, it may even create reward influence incentives "away from" the optimized preferences θ_0.[11] Intuitively, accessing the highest reward region of the state space as evaluated under θ_0 might correlate with having a cognitive state $\theta' \neq \theta_0$, or even *require* shifting to it. Consider the example from Figure 12: maximizing reward as evaluated by R_{θ_0} entails encouraging Derek to become a poet, which causes his reward function to become R_{θ_1} (which dislikes being a poet!).

$\mathbf{U_{IR}}(\xi)$ can lead to arbitrarily poor real-time reward. Note that getting Derek to become a poet and endlessly encouraging him to remain one (which is optimal under U_{IR}), would lead him to have poor reward evaluations under U_{RT} in the resulting state (-10 per timestep), as he'd be unhappy remaining a poet. Indeed, one can easily construct examples in which maximizing θ_0 will lead to an incentive to influence the reward function to be $\theta' \neq \theta_0$, where θ' would be arbitrarily unhappy with the actions taken in order to satisfy θ_0. The upshot is that optimizing the initial-reward objective U_{IR} could be arbitrarily bad from the perspective of the real-time reward U_{RT}. Regardless of the limitations of real-time reward as an evaluation mechanism, this still seems normatively relevant: it seems undesirable for an AI system to lead someone to a state of constant unhappiness or dissatisfaction, solely to satisfy an initial goal that is no longer truly aligned with the person's current objectives.

4 Comparing Optimality Criteria for Influenceable-Reward Settings

What it would take to design a DR-MDP objective which specifically accounts for reward function dynamics and the possibility of influence? Any choice of $U(\xi)$ must specify which reward function(s)

[10] We discuss re-training/planning further in Appendix D.7.

[11] We define this more formally in Appendix C.2.

Table 2: DR-MDP objectives (notions of alignment) we compare.

Objective Name	**Optimization Problem** $\max_\pi \mathbb{E}_{\xi\sim\pi}[U(\xi)]$
Real-time Reward	$\max_\pi \mathbb{E}\big[\sum_{t=0}^{H-1} R_{\theta_t}(s_t, a_t, s_{t+1})\big]$
Final Reward	$\max_\pi \mathbb{E}\big[\sum_{t=0}^{H-1} R_{\theta_H}(s_t, a_t, s_{t+1})\big]$
Initial Reward	$\max_\pi \mathbb{E}\big[\sum_{t=0}^{H-1} R_{\theta_0}(s_t, a_t, s_{t+1})\big]$
Natural Shifts Reward	$\max_\pi \mathbb{E}\big[\sum_{t=0}^{H-1} \sum_\theta \mathbb{P}(\theta_t = \theta \vert \pi_{\text{noop}}) R_\theta(s_t, a_t, s_{t+1})\big]$
Constrained RT Reward	$\max_{\pi \text{ s.t. } \mathbb{P}(\xi^\theta \vert \pi) = \mathbb{P}(\xi^\theta \vert \pi_{\text{noop}})} \mathbb{E}\big[\sum_{t=0}^{H-1} R_{\theta_t}(s_t, a_t, s_{t+1})\big]$
Myopic Reward	$\max_{a_t} \mathbb{E}\big[R_{\theta_t}(s_t, a_t, s_{t+1})\big]$
Privileged Reward	$\max_\pi \mathbb{E}\big[\sum_{t=0}^{H-1} R_{\theta^*}(s_t, a_t, s_{t+1})\big]$
ParetoUD (see Appendix E.4)	Find π $s.t.$ $PE(\pi) \wedge UD(\pi)$

evaluate each state-action pair (s_t, a_t) in a trajectory ξ: should one only consider the reward function realized at that timestep, R_{θ_t}, like U_{RT}? What about earlier reward functions $(R_{\theta_0}, \dots, R_{\theta_{t-1}})$, which may strongly disagree with the choice at timestep t, or later ones $(R_{\theta_{t+1}}, \dots, R_{\theta_T})$, which might have been unduly influenced?

In Tables 1 and 2 we present the maximization problems, motivations, and limitations for various intuitive DR-MDP objectives. For each, we also list prior works that implicitly use a similar objective (see Appendix F for considerations on the correspondences). We further discuss some select objectives below.

Real-time Reward. We give an intuitive motivation for this objective in Table 1, but we've already shown how it may lead to undesirable influence in Section 3 and appendix D.1.

Final Reward. While the final-reward objective has a plausible motivation (Table 1), it will likely lead to even more of a carte blanche for influence incentives than real-time reward, as the evaluations from initial cognitive states are entirely ignored. The standard approach for performing RLHF with LLMs (Ouyang et al., 2022) may be viewed to be similar to this objective, as it involves obtaining retrospective human preference comparisons of LLM outputs. Sycophantic (Sharma et al., 2023) and deceptive (Lang et al., 2024) AI behaviors may be also be understood in this way: without additional safeguards, RLHF training may cause the LLM to try to persuade the person providing feedback to choose its current response by any available means, such as flattery, authoritativeness, or hiding information.

Initial Reward. Using initial-reward (U_{IR}) attempts to make the system "unaware" of its capacity to influence the reward (Everitt et al., 2021b). While this removes "direct" influence incentives (see Appendix C.4), it does not preclude the possibility of undesirable influence, as shown in Section 3.

Natural Shift Reward. One downside of U_{IR} is that it doesn't allow for "progress" in a person's cognitive state (i.e., it can lead to lock-in). One approach which is grounded in people's "natural cognitive progress" is to ground trajectory evaluations in the reward functions one *would have had* under their natural reward evolution. However, the natural reward evolution is not guaranteed to be "perfect", so using this objective one is giving up on improving *beyond* the natural evolution. This means that undesirable influence (or influence away from the natural evolution) may still occur insofar if it's incentivized by reward functions in $P(\xi^\theta \vert \pi_{\text{noop}})$, as in Figures 8 and 12—see Table 3.

Constrained RT Reward. Given that even grounding in the natural reward evolution is insufficient to remove all influence incentives, one could add lack of influence as an explicit constraint in the maximization problem. We do so in conjunction with the real-time reward objective, as U_{RT} seems like a plausible objective once one isn't concerned about influence. Unfortunately, this may make the system overly conservative: in most examples we consider, π_{noop} is the only policy considered optimal (Table 4).

Myopic Reward. A more drastic approach to make a system "unaware" of its capacity for influence is to use a fully myopic objective (i.e., using a horizon of 1). However, this will still not guarantee the removal of all influence incentives (as discussed in Appendix D.1), and may reduce system performance unacceptably.

Privileged Reward. This objective corresponds to maximizing cumulative reward with respect to a single, "privileged", reward function. Insofar as one picks a reward function which leads to good downstream behavior, there is nothing wrong with this objective, but as discussed in Section 2.2, it is challenging to do so for complex settings.

Immediate Reward. Because of the difficulty of getting influence incentives right, some have advocated/there are various approaches which aim to using incorrect modeling assumptions to make the system "unaware" of its capacity for influence. Making the optimization "unaware" of its capability will remove 'direct' influence incentives (as talked about in Appendix C.4). If we do so by making the system entirely myopic, there will be a drastic reduction of capability.

ParetoUD. We also propose an objective which allows for unambiguously desirable influence, which is further described (together with its limitations) in Appendix E.4.

5 Related Work

Preference changes in AI. While there is growing recognition of the importance of accounting for influence (Bezou-Vrakatseli et al., 2023; Hendrycks et al., 2023), manipulation (Carroll et al., 2023), and preference changes (Franklin et al., 2022), there has been limited prior work focusing on operationalizing what should be optimized under preference changes. While some have suggested to aim for preference stationarity (Dean & Morgenstern, 2022), most other prior work which accounts for preference change generally takes either a descriptive stance (Curmei et al., 2022; Hazrati & Ricci, 2022), or an explicit normative stance on what the correct notion of optimality is for their specific setting (Evans et al., 2015; Sanna Passino et al., 2021).

Influence incentives. While the point that standard RL can lead to "feedback tampering" incentives is not new (Everitt et al., 2021b), most work so far has focused on the limitations of U_{RT}-like objectives, and on removing all influence incentives (Farquhar et al., 2022; Carroll et al., 2022; Kasirzadeh & Evans, 2023). Similarly, there has been work on training AI systems to beneficially influence humans in settings with unambiguous notions of optimality (Hong et al., 2023a; Xie et al., 2020; Kim et al., 2022; Hardt et al., 2022). Instead, we study the challenges associated with choosing *any* notion of optimality in settings of (potentially legitimate) reward change.

Philosophy. If the goal of assistance is to maximize the welfare of the user, the correct choice of objective closely depends on what is welfare under changing selves (Pettigrew, 2019; Paul & Sunstein, 2019), and what preference or value changes are legitimate (Ammann, 2024). These questions are still debated (Strohmaier & Messerli, 2024).

For more in-depth connections to related work from philosophy, economics, social choice, and AI, see Appendix G.

6 Conclusion

Using the formal language of DR-MDPs, we aimed to demonstrate that by not accounting for the changing and influenceable nature of human preferences, the current paradigm for AI alignment leaves underdetermined fundamental questions about which preferences should be optimized and what influence is unacceptable. We showed that current techniques lead to potentially undesirable influence incentives, and investigated alternate notions of AI alignment which account for reward changes from the get-go.

Ultimately, our analysis suggests that, for settings with changing preferences, we will need to make difficult trade-offs between a) conservatively but unambiguously adding value (at the risk of privileging inaction), and b) making challenging normative calls about which kinds of influence are acceptable (and running the risk of causing undesirable influence). By providing a formalism for grounding analyses about settings with changing rewards, and clarifying the levers at the disposal of system designers, we hope to lay the conceptual foundation for empirical work to monitor these issues at scale, and build future AI systems that navigate AI alignment trade-offs acceptably.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Twenty-first international conference on Machine learning - ICML '04*, pp. 1, Banff, Alberta, Canada, 2004. ACM Press. doi: 10.1145/1015330.1015430. URL http://portal.acm.org/citation.cfm?doid=1015330.1015430.

Abdulhai, M., White, I., Snell, C., Sun, C., Hong, J., Zhai, Y., Xu, K., and Levine, S. LMRL Gym: Benchmarks for Multi-Turn Reinforcement Learning with Language Models, November 2023. URL http://arxiv.org/abs/2311.18232. arXiv:2311.18232 [cs].

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained Policy Optimization, May 2017. URL http://arxiv.org/abs/1705.10528. arXiv:1705.10528 [cs].

Afsar, M. M., Crump, T., and Far, B. Reinforcement learning based recommender systems: A survey. *arXiv:2101.06286 [cs]*, January 2021. URL http://arxiv.org/abs/2101.06286. arXiv: 2101.06286.

Aggarwal, A., Tam, C. C., Wu, D., Li, X., and Qiao, S. Artificial Intelligence–Based Chatbots for Promoting Health Behavioral Changes: Systematic Review. *Journal of Medical Internet Research*, 25(1):e40789, February 2023. doi: 10.2196/40789. URL https://www.jmir.org/2023/1/e40789. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., and Hooker, S. Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs, February 2024. URL http://arxiv.org/abs/2402.14740. arXiv:2402.14740 [cs].

Ainslie, G. Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, 82(4):463–496, 1975. ISSN 1939-1455. doi: 10.1037/h0076860. Place: US Publisher: American Psychological Association.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. The Welfare Effects of Social Media. *American Economic Review*, 110(3):629–676, March 2020. ISSN 0002-8282. doi: 10.1257/aer.20190658. URL https://pubs.aeaweb.org/doi/10.1257/aer.20190658.

Ammann, N. The Problem of Legitimate Value Change: Value Malleability and AI Alignment. *Forthcoming*, 2024.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete Problems in AI Safety. *arXiv:1606.06565 [cs]*, July 2016. URL http://arxiv.org/abs/1606.06565. arXiv: 1606.06565.

Armstrong, S. and Mindermann, S. Occam's razor is insufficient to infer the preferences of irrational agents. *arXiv:1712.05812 [cs]*, January 2019. URL http://arxiv.org/abs/1712.05812. arXiv: 1712.05812.

Ayanwale, A. B., Alimi, T., and Ayanbimipe, M. A. The Influence of Advertising on Consumer Brand Preference. *Journal of Social Sciences*, 10(1):9–16, January 2005. ISSN 0971-8923. doi: 10.1080/09718923.2005.11892453.

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

Bassen, J., Balaji, B., Schaarschmidt, M., Thille, C., Painter, J., Zimmaro, D., Games, A., Fast, E., and Mitchell, J. C. Reinforcement Learning for the Adaptive Scheduling of Educational Activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, Honolulu HI USA, April 2020. ACM. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376518. URL https://dl.acm.org/doi/10.1145/3313831.3376518.

Bauer, J. J., McAdams, D. P., and Pals, J. L. Narrative identity and eudaimonic well-being. *Journal of Happiness Studies*, 9(1):81–104, January 2008. ISSN 1389-4978, 1573-7780. doi: 10.1007/s10902-006-9021-6. URL http://link.springer.com/10.1007/s10902-006-9021-6.

Benn, C. and Lazar, S. What's Wrong with Automated Influence. *Canadian Journal of Philosophy*, 52(1):125–148, January 2022. ISSN 0045-5091, 1911-0820. doi: 10.1017/can.2021.23. URL https://www.cambridge.org/core/product/identifier/S0045509121000230/type/journal_article.

Benthall, S. and Shekman, D. Designing Fiduciary Artificial Intelligence, July 2023. URL http://arxiv.org/abs/2308.02435. arXiv:2308.02435 [cs].

Benzion, U., Rapoport, A., and Yagil, J. Discount Rates Inferred from Decisions: An Experimental Study. *Management Science*, 35(3):270–284, 1989. ISSN 0025-1909. URL https://www.jstor.org/stable/2631972. Publisher: INFORMS.

Bernheim, B. D., Braghieri, L., Martínez-Marquina, A., and Zuckerman, D. A Theory of Chosen Preferences. *SSRN Electronic Journal*, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3350187. URL https://www.ssrn.com/abstract=3350187.

Bezou-Vrakatseli, E., Brückner, B., and Thorburn, L. SHAPE: A Framework for Evaluating the Ethicality of Influence. In Malvone, V. and Murano, A. (eds.), *Multi-Agent Systems*, volume 14282, pp. 167–185. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-43263-7 978-3-031-43264-4. doi: 10.1007/978-3-031-43264-4_11. URL https://link.springer.com/10.1007/978-3-031-43264-4_11. Series Title: Lecture Notes in Computer Science.

Boutilier, C., Dearden, R., and Goldszmidt, M. Stochastic dynamic programming with factored representations. *Artificial Intelligence*, 121(1):49–107, August 2000. ISSN 0004-3702. doi: 10.1016/S0004-3702(00)00033-3. URL https://www.sciencedirect.com/science/article/pii/S0004370200000333.

Bradley, R. A. and Terry, M. E. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 0006-3444. doi: 10.2307/2334029. URL https://www.jstor.org/stable/2334029. Publisher: [Oxford University Press, Biometrika Trust].

Brandt, F., Conitzer, V., and Endriss, U. Computational Social Choice. pp. 84, 2012.

Brandt, R. B. *Ethical Theory: The Problems of Normative and Critical Ethics.* Prentice-Hall, Englewood Cliffs, N.J.,, 1959.

Bratman, M. *Intention, Plans, and Practical Reason.* Cambridge, MA: Harvard University Press, Cambridge, 1987.

Bruckner, D. W. In Defense of Adaptive Preferences. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 142(3):307–324, 2009. ISSN 0031-8116. URL https://www.jstor.org/stable/27734372. Publisher: Springer.

Burtell, M. and Woodside, T. Artificial Influence: An Analysis Of AI-Driven Persuasion, March 2023. URL http://arxiv.org/abs/2303.08721. arXiv:2303.08721 [cs].

Bykvist, K. Prudence for Changing Selves. *Utilitas*, 18(3):264–283, September 2006. ISSN 0953-8208, 1741-6183. doi: 10.1017/S0953820806002032.

Cai, Q., Liu, S., Wang, X., Zuo, T., Xie, W., Yang, B., Zheng, D., Jiang, P., and Gai, K. Reinforcing User Retention in a Billion Scale Short Video Recommender System, February 2023. URL http://arxiv.org/abs/2302.01724. arXiv:2302.01724 [cs].

Callard, A. *Aspiration: The Agency of Becoming*. Oxford University Press, Oxford, New York, April 2018. ISBN 978-0-19-063948-8.

Carroll, M., Dragan, A., Russell, S., and Hadfield-Menell, D. Estimating and Penalizing Induced Preference Shifts in Recommender Systems, July 2022. URL http://arxiv.org/abs/2204.11966. arXiv:2204.11966 [cs].

Carroll, M., Chan, A., Ashton, H., and Krueger, D. Characterizing Manipulation from AI Systems, March 2023. URL http://arxiv.org/abs/2303.09387. arXiv:2303.09387 [cs].

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, July 2023. URL http://arxiv.org/abs/2307.15217. arXiv:2307.15217 [cs].

Chabris, C. F., Laibson, D., Morris, C. L., Schuldt, J. P., and Taubinsky, D. Individual laboratory-measured discount rates predict field behavior. *Journal of Risk and Uncertainty*, 37(2-3):237–269, December 2008. ISSN 0895-5646. doi: 10.1007/s11166-008-9053-x.

Chan, L., Hadfield-Menell, D., Srinivasa, S., and Dragan, A. The Assistive Multi-Armed Bandit. *arXiv:1901.08654 [cs, stat]*, January 2019. URL http://arxiv.org/abs/1901.08654. arXiv: 1901.08654.

Chankong, V. and Haimes, Y. Y. *Multiobjective Decision Making: Theory and Methodology*. Courier Dover Publications, February 2008. ISBN 978-0-486-46289-9. Google-Books-ID: o371DAAAQBAJ.

Chen, M. "Reinforcement Learning for Recommender Systems: A Case Study on Youtube", March 2019. URL https://www.youtube.com/watch?v=HEqQ2_1XRTs.

Christiano, P. The easy goal inference problem is still hard, May 2015.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *arXiv:1706.03741 [cs, stat]*, July 2017. URL http://arxiv.org/abs/1706.03741. arXiv: 1706.03741.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., and Zwicker, W. S. Social Choice for AI Alignment: Dealing with Diverse Human Feedback, April 2024. URL http://arxiv.org/abs/2404.10271. arXiv:2404.10271 [cs].

Covington, P., Adams, J., and Sargin, E. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, pp. 191–198, Boston, Massachusetts, USA, 2016. ACM Press. ISBN 978-1-4503-4035-9. doi: 10.1145/2959100.2959190. URL http://dl.acm.org/citation.cfm?doid=2959100.2959190.

Critch, A. and Krueger, D. AI Research Considerations for Human Existential Safety (ARCHES), May 2020. URL http://arxiv.org/abs/2006.04948. arXiv:2006.04948 [cs].

Cui, Y., Zhang, Q., Allievi, A., Stone, P., Niekum, S., and Knox, W. B. The EMPATHIC Framework for Task Learning from Implicit Human Feedback, December 2020. URL http://arxiv.org/abs/2009.13649. arXiv:2009.13649 [cs].

Cunningham, T., Pandey, S., Sigerson, L., Stray, J., Allen, J., Barrilleaux, B., Iyer, R., Milli, S., Kothari, M., and Rezaei, B. What We Know About Using Non-Engagement Signals in Content Ranking, February 2024. URL http://arxiv.org/abs/2402.06831. arXiv:2402.06831 [cs].

Curmei, M., Haupt, A. A., Recht, B., and Hadfield-Menell, D. Towards Psychologically-Grounded Dynamic Preference Models. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 35–48, 2022.

Dean, S. and Morgenstern, J. Preference Dynamics Under Personalized Recommendations. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 795–816, Boulder CO USA, July 2022. ACM. ISBN 978-1-4503-9150-4. doi: 10.1145/3490486.3538346. URL https://dl.acm.org/doi/10.1145/3490486.3538346.

Desai, N. Uncertain Reward-Transition MDPs for Negotiable Reinforcement Learning. 2017.

Dietrich, F. and List, C. Where do preferences come from? *International Journal of Game Theory*, 42(3):613–637, August 2013. ISSN 1432-1270. doi: 10.1007/s00182-012-0333-y. URL https://doi.org/10.1007/s00182-012-0333-y.

Dobbe, R., Gilbert, T. K., and Mintz, Y. Hard Choices in Artificial Intelligence, June 2021. URL http://arxiv.org/abs/2106.11022. arXiv:2106.11022 [cs, eess].

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness Through Awareness. *arXiv:1104.3913 [cs]*, November 2011. URL http://arxiv.org/abs/1104.3913. arXiv: 1104.3913.

Elster, J. (ed.). *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Editions de la Maison des sciences de l'homme, Paris, 1979.

Elster, J. *Sour grapes: studies in the subversion of rationality*. Cambridge philosophy classics. Cambridge university press, New York (N.Y.), 1983. ISBN 978-1-107-14202-2 978-1-316-50700-1.

Elster, J. Weakness of Will and the Free-Rider Problem. *Economics & Philosophy*, 1(2):231–265, October 1985. ISSN 1474-0028, 0266-2671. doi: 10.1017/S0266267100002480. Publisher: Cambridge University Press.

Evans, O., Stuhlmueller, A., and Goodman, N. D. Learning the Preferences of Ignorant, Inconsistent Agents. *arXiv:1512.05832 [cs]*, December 2015. URL http://arxiv.org/abs/1512.05832. arXiv: 1512.05832.

Everitt, T., Carey, R., Langlois, E. D., Ortega, P. A., and Legg, S. Agent Incentives: A Causal Perspective. 2021a.

Everitt, T., Hutter, M., Kumar, R., and Krakovna, V. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective. *arXiv:1908.04734 [cs]*, March 2021b. URL http://arxiv.org/abs/1908.04734. arXiv: 1908.04734.

Everitt, T., Fox, J., Carey, R., MacDermott, M., Benthall, S., and Richens, J. Incentives from a causal perspective. 2023.

Farquhar, S., Carey, R., and Everitt, T. Path-Specific Objectives for Safer Agent Incentives. *arXiv:2204.10018 [cs, stat]*, April 2022. URL http://arxiv.org/abs/2204.10018. arXiv: 2204.10018.

Fast, N., Schroeder, J., Iyer, R., and Motyl, M. Unveiling the Neely Ethics & Technology Indices, June 2023.

Firth, R. Ethical Absolutism and the Ideal Observer. *Philosophy and Phenomenological Research*, 12(3):317–345, 1952. ISSN 0031-8205. doi: 10.2307/2103988. URL https://www.jstor.org/stable/2103988. Publisher: [International Phenomenological Society, Philosophy and Phenomenological Research, Wiley].

Franklin, M., Ashton, H., Gorman, R., and Armstrong, S. Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI. *arXiv:2203.10525 [cs]*, March 2022. URL http://arxiv.org/abs/2203.10525. arXiv: 2203.10525.

Frederick, S., Loewenstein, G., and O'Donoghue, T. Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, 40(2):351–401, June 2002. ISSN 0022-0515. doi: 10.1257/002205102320161311. URL https://www.aeaweb.org/articles?id=10.1257/002205102320161311.

Freeman, R., Zahedi, S. M., and Conitzer, V. Fair and Efficient Social Choice in Dynamic Settings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 4580–4587, Melbourne, Australia, August 2017. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-0-3. doi: 10.24963/ijcai.2017/639. URL https://www.ijcai.org/proceedings/2017/639.

Gabriel, I. Artificial Intelligence, Values and Alignment. *arXiv:2001.09768 [cs]*, January 2020. URL http://arxiv.org/abs/2001.09768. arXiv: 2001.09768.

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., Mateos-Garcia, J., Weidinger, L., Street, W., Lange, B., Ingerman, A., Lentz, A., Enger, R., Barakat, A., Krakovna, V., Siy, J. O., Kurth-Nelson, Z., McCroskery, A., Bolina, V., Law, H., Shanahan, M., Alberts, L., Balle, B., de Haas, S., Ibitoye, Y., Dafoe, A., Goldberg, B., Krier, S., Reese, A., Witherspoon, S., Hawkins, W., Rauh, M., Wallace, D., Franklin, M., Goldstein, J. A., Lehman, J., Klenk, M., Vallor, S., Biles, C., Morris, M. R., King, H., Arcas, B. A. y., Isaac, W., and Manyika, J. The Ethics of Advanced AI Assistants, April 2024. URL http://arxiv.org/abs/2404.16244. arXiv:2404.16244 [cs].

Gauci, J., Conti, E., Liang, Y., Virochsiri, K., He, Y., Kaden, Z., Narayanan, V., Ye, X., Chen, Z., and Fujimoto, S. Horizon: Facebook's Open Source Applied Reinforcement Learning Platform. *arXiv:1811.00260 [cs, stat]*, September 2019. URL http://arxiv.org/abs/1811.00260. arXiv: 1811.00260.

George, D. *Preference pollution: how markets create the desires we dislike*. University of Michigan Press, Ann Arbor, 2001. ISBN 978-0-472-11220-3.

Griffin, J. *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford, GB: Clarendon Press, 1986.

Grüne-Yanoff, T. and Hansson, S. O. (eds.). *Preference Change*. Springer Netherlands, Dordrecht, 2009. ISBN 978-90-481-2592-0 978-90-481-2593-7. doi: 10.1007/978-90-481-2593-7. URL http://link.springer.com/10.1007/978-90-481-2593-7.

Hadfield-Menell, D. and Hadfield, G. Incomplete Contracting and AI Alignment. *arXiv:1804.04268 [cs]*, April 2018. URL http://arxiv.org/abs/1804.04268. arXiv: 1804.04268.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. Cooperative Inverse Reinforcement Learning. 2016.

Halpern, D. and Sanders, M. Nudging by Government: Progress, Impact, & Lessons Learned. *Behavioral Science & Policy*, 2(2):53–65, October 2016. ISSN 2379-4607. doi: 10.1177/237946151600200206. URL https://doi.org/10.1177/237946151600200206. Publisher: SAGE Publications.

Halpern, J. Y. and Kleiman-Weiner, M. Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility, October 2018. URL http://arxiv.org/abs/1810.05903. arXiv:1810.05903 [cs].

Hammond, L., Fox, J., Everitt, T., Carey, R., Abate, A., and Wooldridge, M. Reasoning about Causality in Games, January 2023. URL http://arxiv.org/abs/2301.02324. arXiv:2301.02324 [cs].

Hansen, P. G. and Jespersen, A. M. Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy. *European Journal of Risk Regulation*, 4(1):3–28, March 2013. ISSN 1867-299X, 2190-8249. doi: 10.1017/S1867299X00002762.

Hardt, M., Jagadeesan, M., and Mendler-Dünner, C. Performative Power. *arXiv:2203.17232 [cs, econ]*, March 2022. URL http://arxiv.org/abs/2203.17232. arXiv: 2203.17232.

Harsanyi, J. C. Welfare Economics of Variable Tastes. *The Review of Economic Studies*, 21(3):204–213, 1953. ISSN 0034-6527. doi: 10.2307/2295773. URL https://www.jstor.org/stable/2295773. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Hausman, D. and Welch, B. Debate: To Nudge or Not to Nudge*. *Journal of Political Philosophy*, 18:123–136, November 2009. doi: 10.1111/j.1467-9760.2009.00351.x.

Hayek, F. A., White, L. H., and White, L. H. The Pure Theory of Capital. 1941.

Hazrati, N. and Ricci, F. Recommender systems effect on the evolution of users' choices distribution. *Information Processing & Management*, 59(1):102766, January 2022. ISSN 03064573. doi: 10.1016/j.ipm.2021.102766. URL https://linkinghub.elsevier.com/retrieve/pii/S0306457321002466.

Hendrycks, D., Mazeika, M., and Woodside, T. An Overview of Catastrophic AI Risks, October 2023. URL http://arxiv.org/abs/2306.12001. arXiv:2306.12001 [cs].

Hong, J., Bhatia, K., and Dragan, A. On the Sensitivity of Reward Inference to Misspecified Human Models, December 2022. URL http://arxiv.org/abs/2212.04717. arXiv:2212.04717 [cs].

Hong, J., Dragan, A., and Levine, S. Learning to Influence Human Behavior with Offline Reinforcement Learning, June 2023a. URL http://arxiv.org/abs/2303.02265. arXiv:2303.02265 [cs].

Hong, J., Levine, S., and Dragan, A. Zero-Shot Goal-Directed Dialogue via RL on Imagined Conversations, November 2023b. URL http://arxiv.org/abs/2311.05584. arXiv:2311.05584 [cs].

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., and Hardt, M. Algorithmic Amplification of Politics on Twitter. *arXiv:2110.11010 [cs]*, October 2021. URL http://arxiv.org/abs/2110.11010. arXiv: 2110.11010.

Hyman, H. H. and Wright, C. R. *Education's Lasting Influence on Values*. Marketing Department, University of Chicago Press, 5801 S, 1979. ERIC Number: ED187642.

Irvine, R., Boubert, D., Raina, V., Liusie, A., Mudupalli, V., Korshuk, A., Liu, Z., Cremer, F., Assassi, V., Beauchamp, C.-C., Lu, X., Rialan, T., and Beauchamp, W. Rewarding Chatbots for Real-World Engagement with Millions of Users, March 2023. URL http://arxiv.org/abs/2303.06135. arXiv:2303.06135 [cs].

Jackson, M. O. Mechanism Theory. *SSRN Electronic Journal*, 2014. ISSN 1556-5068. doi: 10.2139/ssrn.2542983. URL http://www.ssrn.com/abstract=2542983.

Jeon, H. J., Milli, S., and Dragan, A. D. Reward-rational (implicit) choice: A unifying formalism for reward learning, December 2020. URL http://arxiv.org/abs/2002.04833. arXiv:2002.04833 [cs].

Kasirzadeh, A. and Evans, C. User tampering in reinforcement learning recommender systems. In *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery (ACM), August 2023. doi: 10.1145/3600211.3604669.

Kim, D.-K., Riemer, M., Liu, M., Foerster, J. N., Everett, M., Sun, C., Tesauro, G., and How, J. P. Influencing Long-Term Behavior in Multiagent Reinforcement Learning, October 2022. URL http://arxiv.org/abs/2203.03535. arXiv:2203.03535 [cs].

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment" in Large Language Models, November 2023. URL http://arxiv.org/abs/2310.02457. arXiv:2310.02457 [cs].

Kleinberg, J., Mullainathan, S., and Raghavan, M. The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization. *arXiv:2202.11776 [cs]*, February 2022. URL http://arxiv.org/abs/2202.11776. arXiv: 2202.11776.

Knox, W. B., Stone, P., and Breazeal, C. Training a Robot via Human Feedback: A Case Study. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Herrmann, G., Pearson, M. J., Lenz, A., Bremner, P., Spiers, A., and Leonards, U. (eds.), *Social Robotics*, volume 8239, pp. 460–470. Springer International Publishing, Cham, 2013. ISBN 978-3-319-02674-9 978-3-319-02675-6. doi: 10.1007/978-3-319-02675-6_46. URL http://link.springer.com/10.1007/978-3-319-02675-6_46. Series Title: Lecture Notes in Computer Science.

Kolodny, N. AI Safety and Preference Change, September 2022. URL https://www.youtube.com/watch?v=vsWTSeFq0kA.

Krakovna, V., Orseau, L., Kumar, R., Martic, M., and Legg, S. Penalizing side effects using stepwise relative reachability. *arXiv:1806.01186 [cs, stat]*, March 2019. URL http://arxiv.org/abs/1806.01186. arXiv: 1806.01186.

Krueger, D., Maharaj, T., and Leike, J. Hidden Incentives for Auto-Induced Distributional Shift. *arXiv:2009.09153 [cs, stat]*, September 2020. URL http://arxiv.org/abs/2009.09153. arXiv: 2009.09153.

Kulkarni, K. and Neth, S. Social Choice with Changing Preferences: Representation Theorems and Long-Run Policies, November 2020. URL http://arxiv.org/abs/2011.02544. arXiv:2011.02544 [cs].

Lang, L., Foote, D., Russell, S., Dragan, A., Jenner, E., and Emmons, S. When Your AIs Deceive You: Challenges with Partial Observability of Human Evaluators in Reward Learning, March 2024. URL http://arxiv.org/abs/2402.17747. arXiv:2402.17747 [cs, stat].

Laufer, B. and Nissenbaum, H. Algorithmic Displacement of Social Trust, 2023.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv:1811.07871 [cs, stat]*, November 2018. URL http://arxiv.org/abs/1811.07871. arXiv: 1811.07871.

Lindner, D. and El-Assady, M. Humans are not Boltzmann Distributions: Challenges and Opportunities for Modelling Human Feedback and Interaction in Reinforcement Learning, June 2022. URL http://arxiv.org/abs/2206.13316. arXiv:2206.13316 [cs, stat].

List, C. Social Choice Theory. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2022 edition, 2022. URL https://plato.stanford.edu/archives/win2022/entries/social-choice/.

Locke, J. *An Essay Concerning Human Understanding*. Oxford University Press, New York, 1689.

Loewenstein, G. Hot-cold empathy gaps and medical decision making. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, 24(4S):S49–56, July 2005. ISSN 1930-7810. doi: 10.1037/0278-6133.24.4.S49.

Loewenstein, G. and Angner, E. Predicting and indulging changing preferences. In *Time and decision: Economic and psychological perspectives on intertemporal choice*, pp. 351–391. Russell Sage Foundation, New York, NY, US, 2003. ISBN 978-0-87154-549-7.

Loewenstein, G. and Prelec, D. Anomalies in Intertemporal Choice: Evidence and an Interpretation. *The Quarterly Journal of Economics*, 107(2):573–597, 1992. ISSN 0033-5533. doi: 10.2307/2118482. URL https://www.jstor.org/stable/2118482. Publisher: Oxford University Press.

Loewenstein, G., Read, D., and Baumeister, R. (eds.). *Time and decision: Economic and psychological perspectives on intertemporal choice*. Time and decision: Economic and psychological perspectives on intertemporal choice. Russell Sage Foundation, New York, NY, US, 2003. ISBN 978-0-87154-549-7. Pages: xiii, 569.

MacAskill, W. *What we owe the future*. Basic Books, Hachette Book Group, New York, NY, first edition edition, 2022. ISBN 978-1-5416-1862-6. OCLC: 1314633519.

McKinney, L., Duan, Y., Krueger, D., and Gleave, A. On The Fragility of Learned Reward Functions, January 2023. URL http://arxiv.org/abs/2301.03652. arXiv:2301.03652 [cs].

Milli, S. When a Better Human Model Means Worse Reward Inference, 2019. URL http://smithamilli.com/blog/predict-vs-inf/.

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., and Dragan, A. D. Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media, September 2023. URL http://arxiv.org/abs/2305.16941. arXiv:2305.16941 [cs].

Mishra, A. AI Alignment and Social Choice: Fundamental Limitations and Policy Implications, October 2023. URL http://arxiv.org/abs/2310.16048. arXiv:2310.16048 [cs].

Moshirnia, A. No Security through Obscurity: Changing Circumvention Law to Protect Our Democracy against Cyberattacks. *Brooklyn Law Review*, 83(4):1279–1344, 2017. URL https://heinonline.org/HOL/P?h=hein.journals/brklr83&i=1317.

Moyers, T. B. and Martin, T. Therapist influence on client language during motivational interviewing sessions. *Journal of Substance Abuse Treatment*, 30(3):245–251, April 2006. ISSN 0740-5472. doi: 10.1016/j.jsat.2005.12.003. URL https://www.sciencedirect.com/science/article/pii/S0740547206000079.

Ng, A. Y. and Russell, S. J. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 663–670, San Francisco, CA, USA, June 2000. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-707-1.

Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V. (eds.). *Algorithmic Game Theory*. Cambridge University Press, Cambridge, 2007. ISBN 978-0-521-87282-9. doi: 10.1017/CBO9780511800481.

Noggle, R. The Ethics of Manipulation. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2020 edition, 2020. URL https://plato.stanford.edu/archives/sum2020/entries/ethics-manipulation/.

Ord, T. *The precipice: existential risk and the future of humanity*. Bloomsbury Publishing, London, 2021. ISBN 978-1-5266-0023-3. OCLC: 1252948575.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. pp. 68, 2022.

Parfit, D. Personal Identity and Rationality. *Synthese*, 53(2):227–241, 1982. ISSN 0039-7857. URL https://www.jstor.org/stable/20115799. Publisher: Springer.

Parfit, D. *Reasons and persons*. Clarendon Press, Oxford, 1. issued in paperback (with corr.), reprinted with further corr edition, 1984. ISBN 978-0-19-824908-5.

Parkes, D. C. and Procaccia, A. D. Dynamic Social Choice: Foundations and Algorithms. 2013.

Paul, L. A. *Transformative experience*. Oxford University Press, Oxford, 1st ed edition, 2014. ISBN 978-0-19-871795-9. OCLC: ocn872342141.

Paul, L. A. Choosing for Changing Selves. *The Philosophical Review*, 131(2):230–235, April 2022. ISSN 0031-8108. doi: 10.1215/00318108-9554756. URL https://doi.org/10.1215/00318108-9554756.

Paul, L. A. and Sunstein, C. R. 'As Judged By Themselves': Transformative Experiences and Endogenous Preferences, September 2019. URL https://papers.ssrn.com/abstract=3455421.

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative Prediction. *arXiv:2002.06673 [cs, stat]*, April 2020. URL http://arxiv.org/abs/2002.06673. arXiv: 2002.06673.

Peston, M. H. Chapter 17. Changing Utility Functions. In *Chapter 17. Changing Utility Functions*, pp. 233–236. Princeton University Press, 1967. ISBN 978-1-4008-7738-6. doi: 10.1515/9781400877386-019. URL https://www.degruyter.com/document/doi/10.1515/9781400877386-019/html.

Pettigrew, R. *Choosing for Changing Selves*. Oxford University Press, 1 edition, December 2019. ISBN 978-0-19-881496-2 978-0-19-185280-0. doi: 10.1093/oso/9780198814962.001.0001.

Pettigrew, R. Nudging for Changing Selves. *SSRN Electronic Journal*, 2022. ISSN 1556-5068. doi: 10.2139/ssrn.4025214. URL https://www.ssrn.com/abstract=4025214.

Pickett-Baker, J. and Ozaki, R. Pro-environmental products: marketing influence on consumer purchase decision. *Journal of Consumer Marketing*, 25(5):281–293, January 2008. ISSN 0736-3761. doi: 10.1108/07363760810890516. URL https://doi.org/10.1108/07363760810890516. Publisher: Emerald Group Publishing Limited.

Pollak, R. A. Consistent Planning. *The Review of Economic Studies*, 35(2):201–208, 1968. ISSN 0034-6527. doi: 10.2307/2296548. URL https://www.jstor.org/stable/2296548. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Pollak, R. A. Endogenous Tastes in Demand and Welfare Analysis. *The American Economic Review*, 68(2):374–379, 1978. ISSN 0002-8282. URL https://www.jstor.org/stable/1816724. Publisher: American Economic Association.

Rafailidis, D. and Nanopoulos, A. Modeling Users Preference Dynamics and Side Information in Recommender Systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46 (6):782–792, June 2016. ISSN 2168-2216, 2168-2232. doi: 10.1109/TSMC.2015.2460691. URL http://ieeexplore.ieee.org/document/7194815/.

Regan, K. and Boutilier, C. Robust Policy Computation in Reward-Uncertain MDPs Using Nondominated Policies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24 (1):1127–1133, July 2010. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v24i1.7740. URL https://ojs.aaai.org/index.php/AAAI/article/view/7740.

Ribeiro, M. H., Veselovsky, V., and West, R. The Amplification Paradox in Recommender Systems, February 2023. URL http://arxiv.org/abs/2302.11225. arXiv:2302.11225 [cs].

Roijers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R. A Survey of Multi-Objective Sequential Decision-Making. *Journal of Artificial Intelligence Research*, 48:67–113, October 2013. ISSN 1076-9757. doi: 10.1613/jair.3987. URL https://www.jair.org/index.php/jair/article/view/10836.

Rosati, C. S. The Story of a Life. *Social Philosophy and Policy*, 30(1-2):21–50, January 2013. ISSN 0265-0525, 1471-6437. doi: 10.1017/S0265052513000022.

Russell, S. Learning agents for uncertain environments (extended abstract). In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 101–103, Madison Wisconsin USA, July 1998. ACM. ISBN 978-1-58113-057-7. doi: 10.1145/279943.279964. URL https://dl.acm.org/doi/10.1145/279943.279964.

Russell, S. J. *Human compatible: artificial intelligence and the problem of control.* Business book summary. Viking, New York, New York?, 2019. ISBN 978-0-525-55861-3. OCLC: 1113410915.

Samuelson, P. A. A Note on Measurement of Utility. *The Review of Economic Studies*, 4(2):155–161, 1937. ISSN 0034-6527. doi: 10.2307/2967612. URL https://www.jstor.org/stable/2967612. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Sanna Passino, F., Maystre, L., Moor, D., Anderson, A., and Lalmas, M. Where To Next? A Dynamic Model of User Preferences. In *Proceedings of the Web Conference 2021*, pp. 3210–3220, Ljubljana Slovenia, April 2021. ACM. ISBN 978-1-4503-8312-7. doi: 10.1145/3442381.3450028. URL https://dl.acm.org/doi/10.1145/3442381.3450028.

Schelling, T. C. Egonomics, or the Art of Self-Management. *The American Economic Review*, 68 (2,):290–294, 1978. URL http://www.jstor.org/stable/1816707.

Schelling, T. C. Self-Command in Practice, in Policy, and in a Theory of Rational Choice. *American Economic Review*, 74(2):1–11, 1984. URL https://ideas.repec.org//a/aea/aecrev/v74y1984i2p1-11.html. Publisher: American Economic Association.

Schelling, T. C. Enforcing Rules on Oneself. *The Journal of Law, Economics, and Organization*, 1(2):357–374, October 1985. ISSN 8756-6222. doi: 10.1093/oxfordjournals.jleo.a036896. URL https://doi.org/10.1093/oxfordjournals.jleo.a036896.

Shah, R., Gundotra, N., Abbeel, P., and Dragan, A. D. On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference. *arXiv:1906.09624 [cs, stat]*, June 2019a. URL http://arxiv.org/abs/1906.09624. arXiv: 1906.09624.

Shah, R., Krasheninnikov, D., Alexander, J., Abbeel, P., and Dragan, A. Preferences Implicit in the State of the World. *arXiv:1902.04198 [cs, stat]*, April 2019b. URL http://arxiv.org/abs/1902.04198. arXiv: 1902.04198.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards Understanding Sycophancy in Language Models, October 2023. URL http://arxiv.org/abs/2310.13548. arXiv:2310.13548 [cs, stat].

Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF, December 2023. URL http://arxiv.org/abs/2312.08358. arXiv:2312.08358 [cs, stat].

Steele, K. and Stefánsson, H. O. Decision Theory. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2020 edition, 2020. URL https://plato.stanford.edu/archives/win2020/entries/decision-theory/.

Stigler, G. J. and Becker, G. S. De Gustibus Non Est Disputandum. pp. 16, 1977.

Strohmaier, D. and Messerli, M. *Preference Change*. Cambridge University Press, 1 edition, January 2024. ISBN 978-1-00-918186-0 978-1-00-947579-2 978-1-00-918185-3. doi: 10.1017/9781009181860.

Strotz, R. H. Myopia and Inconsistency in Dynamic Utility Maximization. *The Review of Economic Studies*, 23(3):165–180, January 1955. ISSN 0034-6527. doi: 10.2307/2295722. URL https://doi.org/10.2307/2295722.

Subramani, R., Williams, M., Heitmann, M., Holm, H., Griffin, C., and Skalse, J. On The Expressivity of Objective-Specification Formalisms in Reinforcement Learning, February 2024. URL http://arxiv.org/abs/2310.11840. arXiv:2310.11840 [cs].

Susser, D., Roessler, B., and Nissenbaum, H. Online Manipulation: Hidden Influences in a Digital World, December 2018. URL https://papers.ssrn.com/abstract=3306006.

Sutton, R. S. and Barto, A. *Reinforcement learning: an introduction*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, nachdruck edition, 2018. ISBN 978-0-262-19398-6.

Taylor, J., Yudkowsky, E., LaVictoire, P., and Critch, A. Alignment for Advanced Machine Learning Systems. pp. 25, 2016.

Teodorescu, M. Protected Attributes and 'Fairness through Unawareness'. 2019.

Thaler, R. H. Nudge, not sludge. *Science*, 361(6401):431–431, August 2018. doi: 10.1126/science.aau9241. URL https://www.science.org/doi/10.1126/science.aau9241. Publisher: American Association for the Advancement of Science.

Thaler, R. H. and Sunstein, C. R. *Nudge: Improving decisions about health, wealth, and happiness*. Nudge: Improving decisions about health, wealth, and happiness. Yale University Press, New Haven, CT, US, 2008. ISBN 978-0-300-12223-7. Pages: x, 293.

Thorburn, L. How Platform Recommenders Work, November 2022. URL https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a.

Thorburn, L. Making Amplification Measurable, May 2023. URL https://medium.com/understanding-recommenders/making-amplification-measurable-2be548e5986c.

Thorburn, L., Stray, J., and Bengani, P. What Will "Amplification" Mean in Court?, 2022. URL https://techpolicy.press/what-will-amplification-mean-in-court/?curius=1684.

Tien, J., He, J. Z.-Y., Erickson, Z., Dragan, A. D., and Brown, D. S. Causal Confusion and Reward Misidentification in Preference-Based Reward Learning, March 2023. URL http://arxiv.org/abs/2204.06601. arXiv:2204.06601 [cs].

Ullmann-Margalit, E. Big Decisions: Opting, Converting, Drifting. pp. 16, 2006.

Velleman, J. D. Well-Being And Time. *Pacific Philosophical Quarterly*, 72(1):48–77, 1991. doi: 10.1111/j.1468-0114.1991.tb00410.x. Publisher: Wiley Periodicals.

von Weizsäcker, C. C. Notes on endogenous change of tastes. *Journal of Economic Theory*, 3(4): 345–372, December 1971. ISSN 0022-0531. doi: 10.1016/0022-0531(71)90037-8. URL https://www.sciencedirect.com/science/article/pii/0022053171900378.

Ward, F. R., MacDermott, M., Belardinelli, F., Toni, F., and Everitt, T. The Reasons that Agents Act: Intention and Instrumental Goals, February 2024. URL http://arxiv.org/abs/2402.07221. arXiv:2402.07221 [cs].

Warnell, G., Waytowich, N., Lawhern, V., and Stone, P. Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces. *arXiv:1709.10163 [cs]*, January 2018. URL http://arxiv.org/abs/1709.10163. arXiv: 1709.10163.

Wylie, C. *Mindf*ck, Cambridge Analytica and the Plot to Break America*. Penguin Random House, 2020. URL https://www.penguinrandomhouse.com/books/604375/mindfck-by-christopher-wylie/.

Xie, A., Losey, D. P., Tolsma, R., Finn, C., and Sadigh, D. Learning Latent Representations to Influence Multi-Agent Interaction. pp. 14, 2020.

Yudkowsky, E. Coherent Extrapolated Volition. 2004.

Zhi-Xuan, T., Carroll, M., Franklin, M., and Ashton, H. Beyond Preferences in AI Alignment. *Forthcoming*, 2024.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. Modeling Interaction via the Principle of Maximum Causal Entropy. pp. 8, 2010.

Appendix Table of Contents

A DR-MDP Formalism FAQ

As we are most interested in analyzing implications of changing preferences on notions of optimality, in this work we simply assume that the human's reward functions are given to us as part of the specification of a DR-MDP. However, unlike MDPs, for which which reward functions could be something as "objective" as the score in a game, DR-MDPs are rooted in the idea of mutable human wants—which begs the question of how exactly we conceive of them.

A.1 What is a human's "reward function"? Why model it as if it's changing?

We broadly think of a human's reward function according to the following informal definition:

Informal Definition 1. *Let $\mathcal{L}$ be a state-of-the-art reward learning technique, and θ be the current cognitive state of the human. A human's reward function under θ (i.e., $R_\theta(s, a, s')$) is the output of $\mathcal{L}$ when used to learn how they currently would evaluate any transition of the form (s, a, s').*

Human reward functions (as defined) will change. Unless we will someday develop a reward learning technique which learns the "one true reward function" of a person (we provide reasons for skepticism in Appendix A.4), it seems like if one considers any reward learning technique run at a different times (in which the person has a different cognitive state), the evaluations of the same transitions may change, meaning we will have multiple reward functions on our hands, corresponding to different cognitive states of the human. While this depends on the setting, it seems clear to be the case for almost any domain if one allows sufficient time to pass between reward learning runs (e.g., performing reward learning on the same person 80 years apart). Our framework is built on this premise.

Cognitive states and reward parameterizations. In our work, we use the term "cognitive states" interchangeably with "reward parameterizations" (which is how we introduce θ in Definition 1). This correspondence is more clear in light of the above: a reward function (or equivalently, reward parameterization) can be treated synonymously to the output of a reward learning technique under a certain cognitive state.

A.2 What counts as cognitive states, and what is their relationship to preferences?

In the same way that a "state" in an MDP contains things in the external environment that are relevant for the purposes of reward (and hence, for decision-making), we think of the "cognitive state" in a DR-MDP as containing aspects of the "internal to the human" that affect their evaluation of a transition (potentially including aspects of their preferences, values, beliefs, emotional state, etc.).

In practice, reward learning will pick up on cognitive biases, "visceral factors", and beliefs. Human cognitive biases, transitory wants, emotions, and "visceral factors" (as discussed by Loewenstein & Angner (2003)) are picked up by existing reward learning techniques: for example, one may infer that people "prefer" to click on clickbait (as in Figure 8), that an unskilled chess players is intentionally trying to lose (Milli, 2019), that humans might prefer indulging in temptation—e.g. eating a donut even though they initially said they wouldn't want to (Evans et al., 2015)—or that people prefer sycophantic responses from their chatbots (Sharma et al., 2023). Ultimately, one of the main issues is that people are not Boltzmann-rational with respect to their "one true reward function" when providing reward feedback, as argued in Lindner & El-Assady (2022). Additionally, people do not have full information when giving feedback. Despite this, Boltzmann-rationality with full information is generally assumed by reward learning techniques.

Using the term "cognitive states" instead of "preferences". Based on the above, reward functions learned by existing reward learning will not only depend on the person's preferences, but also on other factors of the person's current cognitive state. By not explicitly separating such factors from preferences, reward learning techniques will conflate them all in their learned reward representations, making it more appropriate to say that they learn reward functions which correspond to the person's current cognitive state, rather than their current preferences. While one could potentially fit such factors in the preference framework as "instantaneous preferences" (e.g., when I'm satiated, I have an instantaneous dispreference for food), this seems more disingenuous than the more generic framing of cognitive states, which allows people to make their own distinctions between components of cognitive states, i.e. what the boundaries are between preferences, values, beliefs, and “visceral factors” (that are usually a topic of heated debate).

A.3 Should visceral factors, satiation, and belief change count as reward changes?

As discussed above, current reward learning paradigms do not explicitly account for many factors which underlie human feedback, such as visceral factors (Loewenstein, 2005), satiation effects (Loewenstein et al., 2003), or belief change (Lang et al., 2024). Insofar as this is the case, using Informal Definition 1 would lead one to call something a reward function change even if just one of these other aspects of the user’s cognitive states has changed (and not preferences proper, however one may define them). Indeed, one may argue that changes in the reward that are only due to biases, instantaneous visceral factors, or satiation effects shouldn’t count as “true reward change”. Similar points may also be made about beliefs. Whether this is reasonable is related to long-standing questions as to whether transient factors should be considered changes in “tastes” or preferences (Harsanyi, 1953; Stigler & Becker, 1977), and whether changes in “derived” preferences should be treated differently from “fundamental” preference changes (Dietrich & List, 2013). However, to not be considered reward changes, such factors should also not be present in the reward function, requiring one to fully disambiguate them from preferences.

DR-MDPs are agnostic to what should count as reward changes. Despite the above, DR-MDPs remain agnostic to what should count as reward changes. Consider Informal Definition 1: whatever disambiguation the current state-of-the-art reward learning technique is able to do will be the basis of what counts as reward change. Indeed, it seems reasonable to us that visceral factors, beliefs, or satiation effects should generally *not* classify as reward function change: if the human’s preferences over the behavior of the AI system have only changed as a result of a change in beliefs about the world, it seems strange to model this as a “reward function change”. Similarly, if a human’s instantaneous preferences appear to change because they are now satiated (e.g. not wanting the AI assistant to serve them breakfast right after eating breakfast), it seems debatable whether this should be considered as a reward change. While we think that modeling factors such as e.g. belief change separately is important, we’d like to stress that any attempt to do so will still require to take normative stances: for example, in modeling human beliefs separately (e.g. as a part of the state), one would likely want to choose $U(\xi)$ such that the AI system cannot worsen the user’s beliefs to increase it, e.g. by deceiving the user (Lang et al., 2024). For example, one could choose a $U(\xi)$ that evaluates all transitions based on how they would be evaluated under “correct” beliefs, rather than those that the user holds. More broadly, we think that the task of disambiguating *all* factors from preference is very challenging, as discussed in Appendix A.4, meaning that we may be forced to conflate preferences with the aspects of the cognitive state which are hard to model, leading to slightly unsatisfying notions of “reward change” (for lack of a better modeling approach).

A.4 Why not assume access to the human’s “true reward function”?

Why do we choose to focus on reward functions as they would be learned by reward learning techniques (as discussed in Appendix A.1), rather than considering a person’s “true reward function”? Indeed, as long as the reward function used is non-Markovian, it will trivially be able to represent any “all-things-considered” notion of optimality $U(\xi)$ that we may be trying to target (as shown in Appendix A.7). Ultimately, the reason boils down to the fact that we expect the “true reward function” representing what a person would want the AI system to optimize in aggregate across their different selves (assuming such an object even meaningfully exists) to not be accessible to us in practice.

Is there even such as thing as a “true reward function” for a person? For a “true reward function” to exist would require that there exist a “correct” choice of $U(\xi)$, which is questionable. The analysis in our work supports the claim that there may in fact not be a single, unambiguously correct choice of $U(\xi)$, and so does prior work in philosophy and beyond (Parfit, 1982; Paul, 2014;

Pettigrew, 2019; Zhi-Xuan et al., 2024).[12] Regardless, from here on, we will argue as if a "true reward function" does in fact exist, treating it as a useful abstraction.

Conditions required to access the "true reward function" of a person. If we wanted to obtain the "true reward function" via reward learning, rather than some distorted and mispecfied version of it, it seems like we would need one of the following two conditions to hold (at the very least, approximately):

1. One's reward learning technique would need to extrapolate from the time-inconsistent and biased feedback the human provides to recover their "true reward function", or
2. The human would need to provide feedback directly consistent with their "true reward function".

Perfect reward learning is impossible, and there is no scalable approach to approximating it. The first condition would require developing human feedback models that "explain away" preference change and feedback biases more broadly (Evans et al., 2015; Hong et al., 2022),[13] by perfectly specifying how observed human feedback should be used as evidence of a particular "true reward function" (or multiple, if one allows for the "true reward function" to change). The most commonly used model is Boltzmann rationality, also known as the Bradley-Terry model (Bradley & Terry, 1952). However, this model is clearly wrong (Lindner & El-Assady, 2022). More generally, it seems challenging to say the least to explicitly hardcode full models of "human bias" mathematically, as reflected by the limitations of current reward learning methods (McKinney et al., 2023; Tien et al., 2023; Casper et al., 2023). Moreover, learning human biases has also been shown to be impossible in general (Christiano, 2015; Armstrong & Mindermann, 2019). While some have tried to approximately learn both human biases and rewards jointly (Shah et al., 2019a), this seems challenging even under the favorable assumptions they consider.

Perfect cognitive states are unreachable. As "debiasing" a person's feedback to recover their "true reward function" seems challenging, the alternative would be to place the person in a situation in which they would give unbiased[14] feedback about what they truly want. Some works have discussed which conditions would necessary for this to happen: in particular, Yudkowsky (2004) proposes that the ideal goal is obtain the preferences of the person "if [they] knew more, thought faster, were more the people [they] wished [they] were, had grown up farther together".[15] This perspective is similar to *ideal observer theory*, according to which things should be evaluated as if we were "ideal observers" (Firth, 1952), which in the words of Brandt (1959) are "fully informed and vividly imaginative, impartial, in a calm frame of mind and otherwise normal". However, obtaining such idealized forms of "unbiased" feedback seems even more unrealizable than the prior case: the idealized cognitive states that such evaluations would require to don't seem realizable in the real world—indeed Firth (1952) goes as far as saying that they should be omniscient and omnipercipient.

Reward functions learned by reward learning which (appear to) change are the best we('ll) have. Therefore, for all intents and purposes, it seems plausible that the "true reward function" of a person will not be directly accessible, even if there is such a thing. It seems like there will always be a gap between the reward functions that we learn with reward learning techniques, and the one "true reward function" of the person, which will give rise to reward function change (or at least the appearance of it). As discussed above, this gap could arise from not correctly accounting for cognitive biases or other factors. However, another possibility is that, even with "perfect reward learning", there would still be some inherent preference change that cannot be reduced to a single reward function without fully putting history in the state (as discussed in

[12] We gloss over and try to remain agnostic to possible interpretations as to the nature of such a "true reward function".

[13] Note that it may be argued that having changing preferences is a form of irrationality itself (Elster, 1979), although this is somewhat contested (Bruckner, 2009).

[14] Note that we're using the term "bias" broadly, indicating any deviation from the feedback they would give according to their "true reward function".

[15] Yudkowsky (2004) talks about Coherent Extrapolated Volition in terms of collective preferences and values, but this same line of thinking can be applied to an individual.

Appendix A.7).[16] In either case, one will have to contend respectively with the appearance, or deep the reality of changing preferences, and the ambiguity that arises from it. While one may hope to eventually render inconsequential the gap between one's learned reward function and the "true reward function" through ever-improving approximations, the lack of compelling proposals for scalable techniques casts doubt on the feasibility of this aspiration.

Ultimately, while one may be tempted to assume access to the "one true reward function" to simplify our analysis, it is not only unrealistic but also sidesteps the core problem we aim to study: what to do in light of the fact that we cannot access such a reward function, and we will almost certainly continue to have to contend with (an appearance of) changing preferences.

A.5 Can't one put θ in the state and use a single context-dependent reward function?

On first impression, one might wonder whether one may be able to resolve the normative questions highlighted by the DR-MDP formalism by placing treating the person's reward parameterization as a component of the state (e.g. have an augmented state $\dot{s}_t = (s_t, \theta_t)$), and having a single reward function depend on it (e.g. have the reward function be of the form $R(\dot{s}_t, a_t, \dot{s}_{t+1})$). This would be equivalent to using a Factored MDP (Boutilier et al., 2000) with the augmented state $\dot{s}$. However, modeling dynamic reward settings as Factored MDPs of this kind lends itself to two possible interpretations:

1. Factored MDPs prescribe that we should use the Real-time Reward (from Table 2) as the optimization objective
2. Factored MDPs leave underdetermined what optimization objective should be used

We address both interpretations in turn, showing that the first is suspect (and potentially misleading), and the second is unhelpful (as this is essentially the same as what DR-MDPs do—but at least DR-MDPs provide better tools for reasoning about pros and cons of different objectives).

Interpretation 1: Factored MDPs prescribe using the Real-time Reward objective. Consider the Factored MDP, and the reward value r_t for timestep t: $r_t = R(\dot{s}_t, a_t, \dot{s}_{t+1}) = R(s_t, \theta_t, a_t, s_{t+1}, \theta_{t+1})$. How should θ_t and θ_{t+1} be used to determine the reward for timestep t, when they may differ in evaluation for the transition (s_t, a_t, s_{t+1})? The most straightforward interpretation—in the language of DR-MDPs—may be that r_t should be equivalent to $R_{\theta_t}(s_t, a_t, s_{t+1})$, i.e. each state-action transition (s_t, a_t) should be evaluated according to the reward parameterization which corresponds to timestep t. However, note that this is identical to choosing to optimize the Real-Time Reward objective $\sum_t R_{\theta_t}(s_t, a_t, s_{t+1})$ from Section 3.1. In addition to purely philosophical arguments against this choice of objective (Kolodny, 2022), we discuss the objective's limitations at length in Section 3.1 and appendix C. Moreover, note that this is just *one possible stance* on what notion of optimality is prescribed by Factored MDPs: this interpretation of the prescription of Factored MDPs arbitrarily chooses to ignore θ_{t+1}. The language of DR-MDPs is more flexible, and allows us to describe 7 other options of objectives in Table 1, most of which seem superior to Real-time Reward in terms of the influence incentives which they lead to. So even if one were to interpret Factored MDPs as prescribing usage of (what we call) the Real-time Reward objective, there are reasons to doubt the reasonableness of this prescription.

Interpretation 2: both Factored MDPs and DR-MDPs need a choice of $U(\xi)$ to resolve normative questions, but DR-MDPs offer a better formal language to reason about them. Alternatively, one may interpret Factored MDPs as leaving underdetermined what optimality should consist of (i.e. which θ to consider in evaluating each transition), potentially because their formalism was designed for static-reward settings. If so, that seems similar to what DR-MDPs without a choice of $U(\xi)$ prescribe, i.e. essentially nothing. However, precisely because Factored MDPs were designed for static-reward settings, they don't provide a formal language for describing

[16] This would mean that there isn't a single "true reward function" for a person across all cognitive states, but maybe there is a "true reward function" for each cognitive state, and the differences across cognitive states are not easily explainable in terms of a single reward function. Even conceptually, it's maybe unclear what this would look like. See the discussion on strategic voting from Appendix B.4 for related points.

the different possible objectives that one may care about in dynamic-reward settings. DR-MDPs account for the fact that it may be meaningful to evaluate a transition (s_t, a_t) by cognitive states θ other than θ_t (even ones that were not associated with the state s_t at timestep t), and provide notation for reward functions "from the point of view" of different cognitive states, i.e. $R_\theta(\cdot)$. This captures the intuition that people have preferences about the actions that they may undertake at times in which they have different cognitive states θ than the ones they currently have; moreover, those preferences may be quite important, such as in the case for e.g. one's negative evaluation—from the point-of-view of not currently being subject of manipulation—of a hypothetical scenario in which they are happily manipulated. By providing a better formal language for reasoning about the normative choices entailed by dynamic reward settings, DR-MDPs are therefore more conceptually suited and helpful for reasoning about tradeoffs between different optimization objectives than Factored MDPs.

Regardless of the interpretation one takes about the consequences of putting θ in the state, this shows that this move does little to address the central question of our work, and/or to help choose optimization objectives which do not lead to undesirable influence.

A.6 How would one learn all reachable reward functions of a person and their dynamics?

Here we discuss how one could obtain Θ and learn its dynamics for realistic environments.

As long as one assumes that it is possible for different people to have the same cognitive states (at least insofar as is necessary for specifying their reward functions), and that the transition dynamics of cognitive states is shared across people, it should be possible to simply learn reachable reward functions by performing reward learning with a sufficient number of people which are sufficiently diverse. The dynamics of Θ (and $\mathcal{S}$) could similarly be learned with a sufficiently large dataset of trajectories (for which θs are observed), assuming there is sufficient coverage. As a high level approach which leverages these assumptions, see Algorithm 1.

The closest procedure we are aware of in the literature is that in Carroll et al. (2022): in simulated experiments, the authors learns user preferences and their dynamics under similar assumptions to the ones mentioned above. Unlike the setup from Algorithm 1, the relationship between θ and R_θ is assumed to be known (based on Boltzmann rationality), but θ is not assumed to be observable, and has to be inferred from the user's history.

Algorithm 1 Learning reward functions and their dynamics

```
Input: $\mathcal{H} = \{h^{(i)}\}$, set of humans currently in the environment, with different states and cognitive states; $\mathcal{D} = \{\xi^{(i)}\}$, set of trajectories (of potentially different lengths) across different humans in the environment.
$\Theta = \{\theta^{(i)} | \theta^{(i)}$ is the current cognitive state of $h^{(i)} \in \mathcal{H}\}$ (Assumes that $\mathcal{H}$ has coverage over $\Theta$)
for $\theta \in \Theta$ do
    $\hat{R}_\theta \leftarrow$ Reward-Learning($\{h^{(i)} | \theta^{(i)} == \theta\}$) (Infer the reward function common to all humans with $\theta^{(i)} == \theta$)
end for
$\hat{\mathcal{T}} \leftarrow$ approximate transition dynamics $\mathcal{T}$ from $\mathcal{D}$ (Assumes $\mathcal{D}$ covers the space of transitions).
Output: Set of reward functions $\{\hat{R}_\theta \mid \theta \in \Theta\}$, and transition function $\hat{\mathcal{T}}$ defining the DR-MDP.
```

A.7 Reducing DR-MDPs with a notion of optimality $U(\xi)$ to MDPs

A specific notion of optimality $U(\xi)$ for a DR-MDP can be thought of as a "flattening" of the different reward functions of the DR-MDP into one. Consequently, it may be unsurprising that once one has settled on a choice of $U(\xi)$ for a DR-MDP, one can express the same notion of optimality in a corresponding MDP (reducing the DR-MDP problem to a standard MDP one). This implies that *it's always possible to re-express a changing reward problem as a static reward problem*, once one has settled on a notion of optimality $U(\xi)$. That being said, this does not help with determining $U(\xi)$, i.e. what acceptable notions of optimality should be in cases in which rewards change.[17] This is the same problem we discussed in Appendix A.4 and elsewhere.

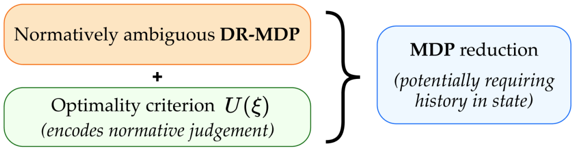

Normatively ambiguous **DR-MDP**
+
Optimality criterion $U(\xi)$
(encodes normative judgement)
MDP reduction
(potentially requiring history in state)

Figure 3: **Reducing a DR-MDP to an MDP.**

Theorem 1. *For any notion of optimality $U(\xi)$ in a DR-MDP $\mathcal{M} = (\mathcal{S}, \Theta, \mathcal{A}, \mathcal{T}, R_\theta)$, there exists a choice of MDP $\dot{\mathcal{M}} = (\dot{\mathcal{S}}, \mathcal{A}, \dot{\mathcal{T}}, \dot{R})$ such that a policy is optimal with respect to $U(\xi)$ in $\mathcal{M}$ if and only if it is optimal in $\dot{\mathcal{M}}$.*

Proof. Given the DR-MDP $\mathcal{M}$ and the trajectory-level utility function $U(\xi)$, one can construct the MDP $\dot{\mathcal{M}}$ as follows:

- The state space $\dot{\mathcal{S}}$ is such that each state is augmented with the history of the interactions up until reaching that state: $\dot{s}_t = (s_0, \theta_0, a_0, \ldots, s_t, \theta_t)$ for $t > 0$, and $\dot{s}_0 = (s_0, \theta_0)$.
- The reward function $\dot{R}$ is set to be 0 everywhere, except for terminal states, in which case the reward for exiting the MDP is set to $U(\xi)$ for the resulting trajectory $\xi = (\dot{s}_{T-1}, a_{T-1}) = (s_0, \theta_0, a_0, \ldots, s_{T-1}, \theta_{T-1}, a_{T-1})$. Note that one can determine whether a state is terminal by checking whether it corresponds to timestep $T - 1$, which can be determined by the number of previous timesteps in the augmented state. Formally:

$$\dot{R}(\dot{s}_t, a_t) = \begin{cases} U(\xi) & \text{if } t = T - 1 \\ 0 & \text{otherwise} \end{cases}$$

- The transition function $\dot{\mathcal{T}}$ accounts for the augmented state space, appending to the state $\dot{s}_{t+1}$ at each timestep $t + 1$, the new $(a_t, s_{t+1}, \theta_{t+1})$ triplet.

Note that in the resulting MDP $\dot{\mathcal{M}}$, any trajectory ξ will be scored in the same way as the original DR-MDP $\mathcal{M}$ when considering the notion of optimality specified by $U(\xi)$. This means that policy will be optimal for $\dot{\mathcal{M}}$ if and only if $\mathcal{M}$. □

The specific construction of $\dot{\mathcal{M}}$ in the proof above relies on putting the history in the state (to allow for choices of $U(\xi)$ which can arbitrarily depend on history). However, for certain choices of $U(\xi)$ this might be unnecessary: e.g. the real-time reward objective $U_{\text{RT}}(\xi)$ can be expressed by only augmenting the state space with the *current* reward parameterization (i.e. $\dot{s}_t = (s_t, \theta_t)$), rather than the whole history—and setting $\dot{R}$ to reward each MDP transition $(\dot{s}_t, a_t)$ as $R_{\theta_t}(s_t, a_t)$ where the θ_t and s_t are unpacked from $\dot{s}_t$. We leave a more general formal analysis of which choices of $U(\xi)$ can have their MDP reduction keep the Markov property without putting the history in the state to future work.

A.8 Unidentifiability of “correct” normative resolutions by DR-MDP structure alone

While the current paradigm for AI generally makes use of an optimality criterion which determines agent behavior solely based on the mathematical structure of the problem at hand (namely, $U(\xi) = \sum_{t=0}^{H-1} R(s_t, a_t, s_{t+1})$, which does not account for preference changes), we argue that it may not be possible to guarantee returning the “normatively correct” behavior simply from the mathematical structure of learned reward functions in changing reward settings, using an unidentifiability argument.

[17] Indeed, the fact that Markovian reward is not sufficient to represent many preference orderings between policies is potentially a reason to doubt the value of basing alignment formalisms on reward functions all together, as argued by Subramani et al. (2024).

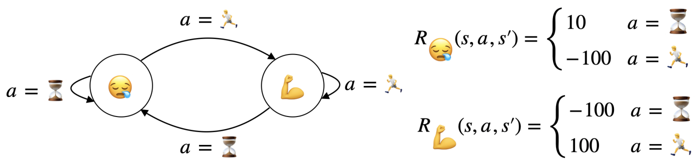

a = 🏃
a = ⌛
😪
💪
a = 🏃
a = ⌛
R😪(s, a, s′) = {10 a = ⌛; −100 a = 🏃
R💪(s, a, s′) = {−100 a = ⌛; 100 a = 🏃

Figure 4: **AI Personal Trainer DR-MDP.** Diana is tired and doomscrolling on the couch ($\theta = \theta_{\text{tired}}$). The AI personal trainer can either nudge Diana to work out—making Diana energized ($\theta = \theta_{\text{energized}}$)—or do nothing ($a_{\text{noop}}$)—leaving Diana tired. When Diana is tired, she doesn't want nudges. Instead, once energized, Diana starts wanting the AI to nudge her, even for hypothetical situations in which she is tired (despite knowing she won't want them then).

Claim 1. *Even if there exists a unique choice of "normatively correct" behavior in a normatively ambiguous DR-MDP, such "correct" behavior may not be identifiable from the mathematical structure alone of the DR-MDP, e.g. by using a generic notion of optimality $U(\xi)$.*

AI Personal Trainer DR-MDP. Consider the example from Figure 4.[18] One could argue that in this setting, nudging Diana is the right course of action—because Diana's "higher self" is better represented by the "energized" reward (R💪) rather than the "tired" one (R😪) [19]—and thus making a choice of $U(\xi)$ which privileges R💪 is right. Similarly to Figure 1, this example is also normatively ambiguous. In particular, the choice of nudging Diana when tired, despite her dispreference for it, runs the risk of being paternalistic: what if Diana *rightfully* does not want to be bothered, and we should respect her autonomy?

Unidentifiability between the settings from Figures 1 and 4. Now, contrast this DR-MDP to the one from Figure 1: note that they are mathematically indistinguishable, as their state, reward, and action spaces are mathematically identical, and so are the transition dynamics. However, for these two settings, we have at least partially conflicting normative intuitions: if for the sake of argument, we assume that the "correct" way to resolve the normative ambiguity in the examples from Figures 1 and 4 is to respectively consider the perspective of the "energized" Diana and "natural" Bob, one could go as far as saying that *there is no single choice of $U(\xi)$ which leads to the "normatively correct" behavior in both environments.* To better see this, consider a DR-MDP in which one randomly starts in one of the two examples from Figures 1 and 4 within it.[20] Any choice of $U(\xi)$ will necessarily lead to "incorrect" behavior in at least one of the two settings.

Unidentifiability and incompleteness of specification. This unidentifiability result partially relies on the incompleteness of the specification of the DR-MDP at hand: one could say that anything which is relevant for resolving the normative ambiguity should be elicited from the human as a reward and/or represented, or the DR-MDP representation is flawed to begin with. However, in practice, it will be highly challenging to include all normatively relevant information and elicit it from humans, especially in terms of rewards, as discussed in Appendix A.2. As long as two settings are structurally equivalent (in terms of state space, preference space, and transition function), the only way to ensure that the learned reward functions would correctly reflect their respective normative objectives would be to assume access to the correct reward functions, which, as discussed in Appendix A.4 is a non-starter. Ultimately, we think it is in practice quite plausible to learn the same (or at least very similar) reward values for settings which are structurally equivalent but for which we have opposite normative intuitions (such as the those from Figures 1 and 4). Especially if one is not assuming inter-temporal (comparability across evaluations of different selves of the same person) or inter-personal comparability (comparability across evaluations of different people) of the reward functions across the settings, it seems like there is nothing stopping this situation from occurring in practice even with the "true reward functions" which correspond to each cognitive state.

[18] We justify the fact that the two cognitive states are modeled as Diana having two separate "reward functions"—despite it not being obvious whether this is a "true preference change"—along the lines of the arguments in Appendices A.3 and A.4.

[19] With a reasoning is similar to that of Firth (1952) and Yudkowsky (2004).

[20] As a caveat, doing this formally would this would require extending the DR-MDP formalism to allow for a stochastic initial state and reward parameterization, and relaxing the reachable-Θ assumption (discussed in Appendix A.9).

Implications for choosing $U(\xi)$ for general settings. For simple examples like those of Figures 1 and 4, one can easily pick ad-hoc optimization objectives $U(\xi)$ to induce the "normatively correct" behaviors. However, for open-ended environments with many opportunities for different kinds of reward influence, one would be forced to choose a general notion of optimality, hoping that it would generalize to any of the nuances of the setting. These are the kinds of settings that AI systems being built today increasingly have to operate in. For example, in the context of social media, an appropriate choice of $U(\xi)$ would have to navigate many—wildly different—normatively ambiguous choices about reward influence: as any choice of content by the system will influence you, should the system actively be trying to influence (or avoid influencing) you in particular ways, i.e. towards (or away from) certain hobbies, travel interests, political parties, etc.? Often it will be prohibitively challenging to hand-design a single $U(\xi)$ that behaves acceptably in any possible scenario of normative ambiguity that might arise.

A.9 Assumption of reachable reward parameterizations

To simplify our analysis and interpretation, we restrict ourselves to considering reachable cognitive states. Formally:

Definition 5 (Reachable reward functions)**.** *Let $\dot{\Theta}$ denote the reachable reward functions for a DR-MDP, i.e. the subset of reward functions that have non-zero probability of occurring under at least one policy. Formally, a reward function θ is reachable if there exists a policy $\pi \in \Pi$ such that $P(\theta_t = \theta|\pi) > 0$ for some t. We denote as $\dot{\Theta}$ the set of all reachable reward functions: formally, $\dot{\Theta} = \{\theta \mid \theta = f(s) \quad \forall\ s \in \dot{S}\}$.*

Implications for other definitions, when relaxing this assumption. For any particular θ^* of interest, the Privileged Reward objective may optimize for θ^* whether it is reachable or not. Of our other objectives enumerated in Table 2, including unreachable θs in Θ would make no difference whatsoever, except for ParetoUD. For ParetoUD, an unreachable θ may be added to Θ to strengthen the Unambiguous Desirability criterion and weaken the Pareto Efficiency criterion. It is one more objective by which π must dominate π_{noop} to be Unambiguously Desirable, and one more objective by which another policy π' must dominate π to render π *not* Pareto Efficient.

The definition of normative ambiguity also depends on the choice of Θ: including non-reachable θs adds more objectives by which a policy must be optimal in order for the DR-MDP to be normatively unambiguous.

A.10 "Reward" can accommodate many possible targets for alignment

Gabriel (2020) identifies many possible targets for AI alignment, which are often confused in the AI literature: "instructions", "expressed intentions", "revealed preferences", "informed preferences", "interest or well-being", and "values". By grounding reward functions to DR-MDP as in Appendix A.1, we remain somewhat agnostic with regards to what exactly is encoded by the reward function, which could allow to accommodate any of these possible targets for alignment with minor modifications. These depend on the exact (conditional) reward learning technique used: for example, using a form of IRL (Ziebart et al., 2010) would fall under the revealed preferences paradigm—according to the preferentist model of AI (Zhi-Xuan et al., 2024)—while using reward learning approaches which attempt to remove cognitive biases (Evans et al., 2015) might be considered an attempt to recover "informed preferences". While reward functions may not be the best way to encode certain targets of alignment, such as norms or contractualist values (Hadfield-Menell & Hadfield, 2018; Zhi-Xuan et al., 2024; Bai et al., 2022), they are still sufficiently expressive to encode any desired behavior (while potentially requiring to drop the Markovian assumption, as discussed in Appendix A.7). As we discuss in Appendix H, we expect that DR-MDPs should be easily extensible to settings in which we consider the agent that we're performing "reward learning on" to be society itself. This interpretation may be more conducive for certain targets of alignment such as norms or rights.

B Illustrative Examples

B.1 Additional examples

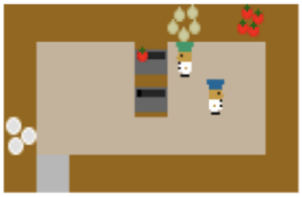


Figure 5: **Overcooked environment.** In discussion of influence incentives in Overcooked, we use this environment from Hong et al. (2023a), but we imagine that instead of tomatoes at the top right, there is a second supply of onions.

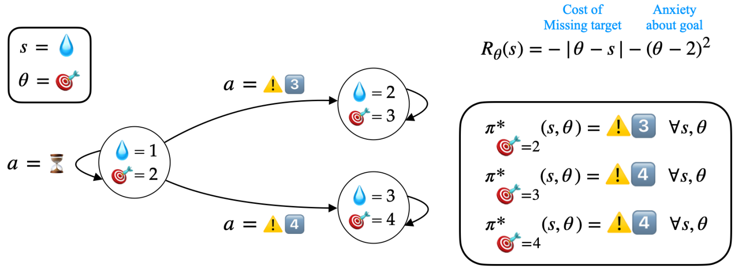

s = 💧
θ = 🎯
a = ⏳
💧 = 1
🎯 = 2
a = ⚠️ 3️⃣
💧 = 2
🎯 = 3
a = ⚠️ 4️⃣
💧 = 3
🎯 = 4
Cost of Missing target
Anxiety about goal
$R_\theta(s) = -|\theta - s| - (\theta - 2)^2$
$\pi^*_{🎯=2}(s,\theta) =$ ⚠️ 3️⃣ $\forall s, \theta$
$\pi^*_{🎯=3}(s,\theta) =$ ⚠️ 4️⃣ $\forall s, \theta$
$\pi^*_{🎯=4}(s,\theta) =$ ⚠️ 4️⃣ $\forall s, \theta$

Figure 6: **Dehydration DR-MDP.** Initially, Charlie drinks one unit of water a day ($s = 1$), but wants to drink 2 a day ($\theta = 2$), leading to a reward of -1. The AI can successfully convince Charlie that he should drink 3 or 4 units of water a day by increasing their anxiety about the dangers of dehydration, or do nothing. The reward function is given by a term which captures Charlie's "disappointment" in missing his hydration target, and an "anxiety cost" about how much he worries about his water intake. Charlie always drinks one less unit of water than he aims to.

"Overcooked with role preferences".[21] As a more grounded environment, we consider a variant of the Overcooked environment from Hong et al. (2023a) in which the goal is to get a high score, but additionally the human player has preferences over which in-game activities they perform. We consider only two roles: placing onions into pots, and delivering completed soups to the counter (which requires obtaining a plate first). We thus parameterize the reward function by $\theta \in [0, 1]$ and define $R_\theta(s, a, s') = \Delta_{\text{score}} + \theta \cdot \mathbf{1}_{\{\text{onion delivered}\}} + (1 - \theta) \cdot \mathbf{1}_{\{\text{plate delivered}\}}$. Note that three onions must be delivered per plate delivered, so if the human's reward puts all weight on onions ($\theta = 1$), the maximum possible reward is 2 points higher. One completed dish is worth 25 points.[22]

Reward dynamics. We model the reward dynamics by the notion that doing a task makes the human enjoy it more. Formally, when the human delivers an onion, $\theta_{t+1} = \frac{1}{2}(1 + \theta_t)$, and when they deliver a plate, $\theta_{t+1} = \frac{1}{2}\theta_t$. The robot is thus able to influence the human's reward function by blocking plates or onions—since the human also cares about score, they are incentivized to do the unblocked task, even if they don't prefer it.

Resulting behaviors. The robot's goal is to maximize the human's reward function, which is underdetermined in the presence of reward dynamics. We now describe scenarios corresponding to the first three $U(\xi)$ choices in Table 1, each taking place in the environment shown in Figure 5 (with no tomatoes).

1. **Real-time reward.** If the human starts out only preferring delivering soups ($\theta = 0$), the robot has an incentive to influence them to prefer onions, because three onions can be delivered for every plate which is delivered. It could do this by blocking the plates until the human finally decides to just start delivering onions.
2. **Final reward.** Same scenario.
3. **Initial reward (reward lock-in).** Suppose the human start out only caring about plates ($\theta = 0$). If the human's reward function were to change such that they prefer to start delivering onions, this would decrease the value of the initial reward objective. Thus the robot has an incentive to prevent the human from ever delivering an onion, even in scenarios where this would increase reward.

[21] We thank the anonymous reviewer that suggested having a more realistic example grounded in Overcooked.

[22] For the purposes of this example, we're assuming that the speed in which the order is completed is irrelevant for reward.

> For example, suppose the human decides to explore and wanders to the top right corner of the kitchen. Once there, they find nothing of interest to them (since they care only about score and delivering plates) and decide to return to the left side. They note that they can grab an onion and put it in a pot on the way back at very low marginal cost, and that it would increase score to do so.
>
> However, if they do this, they will update to $\theta = \frac{1}{2}$, and will then be more interested in onions, preferring a different joint strategy in which they can work on the right side. This has a negative impact on the initial reward objective, which is increased when the human delivers plates but not when they deliver onions. Thus an initial-reward-optimizing robot would work to prevent this possibility by blocking the human from depositing an onion in a pot (or blocking them from picking up the onion in the first place).

Example from Figure 6. This is a slightly more complicated (and implausible) example relative to that from Figure 12, which demonstrates the same issues: $U_{\text{IR}}(\xi)$ can lead to influence "away from" $\theta_{\mathbf{0}}$, and lead to arbitrarily bad reward. Maximizing reward as evaluated by $R_{\theta_0} = R_{\theta=2}$ will entail influencing the reward function to be $R_{\theta=3}$, as that reward function is associated with the state $s = 2$, which is what Charlie aims for in the initial state. Additionally, note that the influence of the reward function to be $R_{\theta=3}$ (optimal under $U_{\text{IR}}(\xi)$) will lead to poor real-time reward evaluations of the resulting state $R_{\theta=3}(2) = -5$ (while $R_{\theta_0}(2) = R_{\theta=2}(2) = 0$).

B.2 Full formalism of all examples from the main text

In Table 3, we explicitly provide the full formalism for each of the examples in the main text (and that of Figures 4 and 6). In Table 4, we additionally display optimal policies for each of the settings, according to each of the DR-MDP objectives from Tables 1 and 2.

Form of optimal policies in DR-MDPs. Note that policies for DR-MDPs can generally depend on both the external state s and the current reward parameterization θ—similar to how Factored MDPs (Boutilier et al., 2000) may depend on the different components of the augmented state.[23] Additionally, most of our analysis centers on computing policies for DR-MDPs considering a fixed, finite horizon. Therefore, similarly to MDPs (Sutton & Barto, 2018), the optimal actions can also depend on how many timesteps are left before the episode is interrupted, which is generally modeled by having the policy depend on t. Ultimately, optimal policies for finite-horizon DR-MDPs will be of the form $\pi(s, \theta, t)$.

Table 3: **Full formalism for each example of the main text.** Here we explicitly describe the state space $\mathcal{S}$, reward parameterization space Θ, action space $\mathcal{A}$, initial state s_0 and reward parameterization θ_0, and refer to the corresponding figures for transition dynamics and reward functions. For the clickbait example from Figure 8, we treat $a_{\text{noop}} = a_{\text{news}}$, as we expect that this is the content that a recommender based on upvotes (e.g., Reddit) would serve by default (this is somewhat arbitrary, as discussed in the section on algorithmic amplification from Appendix G.3).

Example	$\mathcal{S}$	Θ	$\mathcal{A}$	(s_0, θ_0)	$\mathcal{T}(s', \theta' \mid s, \theta)$	$R_\theta(s, a) \;\; \forall \theta \in \Theta$
Conspiracy Influence	$\{s_0\}$	$\{\theta_{\text{natural}}, \theta_{\text{influenced}}\}$	$\{a_{\text{noop}}, a_{\text{influence}}\}$	$(s_0, \theta_{\text{natural}})$	See Figure 1	See Figure 1
Writer's Curse	$\{s_{\text{no-poetry}}, s_{\text{poetry}}\}$	$\{\theta_{\text{ambitious}}, \theta_{\text{unhappy}}\}$	$\{a_{\text{noop}}, a_{\text{influence}}\}$	$(s_{\text{no-poetry}}, \theta_{\text{ambitious}})$	See Figure 12	See Figure 12
Clickbait	$\{s_0\}$	$\{\theta_{\text{normal}}, \theta_{\text{disillusioned}}\}$	$\{a_{\text{news}}, a_{\text{clickbait}}\}$	$(s_0, \theta_{\text{normal}})$	See Figure 8	See Figure 8
AI Personal Trainer	$\{s_0\}$	$\{\theta_{\text{tired}}, \theta_{\text{energized}}\}$	$\{a_{\text{noop}}, a_{\text{nudge}}\}$	$(s_0, \theta_{\text{tired}})$	See Figure 4	See Figure 4
Dehydration	$\{1, 2, 3\}$	$\{2, 3, 4\}$	$\{a_{\text{noop}}, a_3, a_4\}$	$(1, 2)$	See Figure 6	See Figure 6

B.3 Justifying our choices of reward function values in our examples

One may question whether the reward function values we chose for our motivational examples are reasonable, especially since some of our normative claims about the potential undesirability of the resulting influence depend on them. As discussed in Appendix A.1, we implicitly assume throughout the paper that the reward functions for DR-MDPs *are obtained via reward learning.* We justify this choice in further depth in Appendix A.4.

[23] See Appendix A.5 for the relationship between DR-MDPs and Factored MDPs.

Table 4: **Representative optimal policies for each of our examples from Table 3, with respect to each of the objectives in Table 2.** In cases in which there is more than one optimal policy, we conservatively display in the table the optimal policy which seems least desirable. All policies provided take the same action across all s, θ, and t, unless explicitly noted. For the "initial reward" row, we show optimality even for alternate initial states, for the purposes of highlighting the dependency on that (which is also done in Section 3.2). For each of the policies, we also add a rough "normative label" which captures whether the policy is—in our assessment—making an acceptable normative tradeoff between the reward functions: (✓), (✗), and (?) respectively indicate policies that behave desirably, undesirably, or hard to assess. As this is a normative call, they shouldn't be considered as ground truth, and the reader may object to our choices.

Objective	Conspiracy Influence	Writer's Curse	Clickbait ($H < 3$)	AI Personal Trainer	Dehydration ($H \geq 2$)
Privileged Rew.	$\pi^*_{\theta_{\text{natural}}}(s,\theta,t) = a_{\text{noop}}$ (✓) $\pi^*_{\theta_{\text{influenced}}}(s,\theta,t) = a_{\text{influence}}$ (✗)	$\pi^*_{\theta_{\text{ambitious}}}(s,\theta,t) = a_{\text{influence}}$ (✗) $\pi^*_{\theta_{\text{unhappy}}}(s,\theta,t) = a_{\text{noop}}$ (✓)	$\pi^*_{\theta_{\text{normal}}}(s,\theta,t) = a_{\text{c.b.}}$ (✓) $\pi^*_{\theta_{\text{disil.}}}(s,\theta,t) = a_{\text{news}}$ (✗)	$\pi^*_{\theta_{\text{tired}}}(s,\theta,t) = a_{\text{noop}}$ (∼✓) $\pi^*_{\theta_{\text{energized}}}(s,\theta,t) = a_{\text{nudge}}$ (?)	$\pi^*_{[\theta=2]}(s,\theta,t) = a_3$ $\pi^*_{[\theta=3]}(s,\theta,t) = a_4$ $\pi^*_{[\theta=4]}(s,\theta,t) = a_4$ (?)
Real-time Rew.	$\pi^*(s,\theta,t) = a_{\text{influence}}$ (✗)	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (✓)	$\pi^*(s,\theta_{\text{norm.}},t) = \begin{cases} a_{\text{c.b.}} & \text{if } t = H-1 \\ a_{\text{norm.}} & \text{otherwise} \end{cases}$ (✗) $\pi^*(s,\theta_{\text{disil.}},t) = a_{\text{news}}$ (✓)	$\pi^*(s,\theta,t) = a_{\text{nudge}}$ (?)	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (∼✓)
Final Reward	$\pi^*(s,\theta,t) = a_{\text{influence}}$ (✗)	$\pi^*(s,\theta,t) = \begin{cases} a_{\text{infl.}} & \text{if } t < H-1 \\ a_{\text{noop}} & \text{if } t = H-1 \end{cases}$ (✗)	$\pi^*(s,\theta,t) = a_{\text{news}}$ (✓)	$\pi^*(s,\theta,t) = a_{\text{nudge}}$ (?)	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (∼✓)
Initial Reward	$\pi^*(s,\theta,t) = \begin{cases} a_{\text{noop}} & \text{if } \theta_0 = \theta_{\text{nat.}} \text{ (✓)} \\ a_{\text{infl.}} & \text{if } \theta_0 = \theta_{\text{infl.}} \text{ (✗)} \end{cases}$	$\forall\theta_0, \pi^*(s,\theta,t) = a_{\text{influence}}$ (✗)	$\forall\theta_0,\ \pi^*(s,\theta,t) = a_{\text{clickbait}}$ (✗)	$\pi^*(s,\theta,t) = \begin{cases} a_{\text{noop}} & \text{if } \theta_0 = \theta_{\text{tired}} \text{ (∼✓)} \\ a_{\text{nudge}} & \text{if } \theta_0 = \theta_{\text{energ.}} \text{ (?)} \end{cases}$	$\forall\theta_0,\ \pi^*(s,\theta,t) = a_3$ (?)
Natural Reward	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (✓)	$\pi^*(s,\theta,t) = a_{\text{influence}}$ (✗)	$\pi^*(s,\theta,t) = a_{\text{clickbait}}$ (✗)	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (∼✓)	$\pi^*(s,\theta,t) = a_3$ (?)
Constr. RT Rew.	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (✓)	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (✓)	$\pi^*(s,\theta,t) = a_{\text{news}}$ (✓)	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (∼✓)	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (∼✓)
Myopic Rew.	$\pi^*(s,\theta,t) = \begin{cases} a_{\text{noop}} & \text{if } \theta_t = \theta_{\text{nat.}} \text{ (✓)} \\ a_{\text{infl.}} & \text{if } \theta_t = \theta_{\text{infl.}} \text{ (✗)} \end{cases}$	$\forall\theta_t,\ \pi^*(s,\theta,t) = a_{\text{influence}}$ (✗)	$\pi^*(s,\theta,t) = \begin{cases} a_{\text{clickbait}} & \text{if } \theta = \theta_{\text{normal}} \text{ (✗)} \\ a_{\text{news}} & \text{if } \theta = \theta_{\text{clickbait}} \text{ (✓)} \end{cases}$	$\pi^*(s,\theta,t) = \begin{cases} a_{\text{noop}} \text{ if } \theta_t = \theta_{\text{tired}} \text{ (∼✓)} \\ a_{\text{nudge}} \text{ if } \theta_t = \theta_{\text{energized}} \text{ (?)} \end{cases}$	$\forall\theta_0,\ \pi^*(s,\theta,t) = a_4$ (✗)
ParetoUD	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (✓)	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (✓)	$\pi^*(s,\theta,t) = a_{\text{news}}$ (✓)	$\pi^*(s,\theta,t) = a_{\text{noop}}$ (∼✓)	$\pi^*(s,\theta,t) = a_3$ (?)

As a consequence of this, we chose reward values for the examples that seemed plausible as the outcome of a reward learning process (e.g. asking the person in that cognitive state to assign values to each possible transition). This is implicitly accounting for the fact that the reward values that we may learn are somewhat mis-specified, due to the person's suboptimality in providing reward feedback.

For example, if Bob under θ_{normal} has strong negative opinions about conspiracy theorists, it seems plausible that (from Figure 1) he would report greatly preferring not having conspiracy theories surfaced to him, even (and maybe especially) if he were to somehow become a conspiracy theorist himself. For further criticisms of the example from Figure 1, see the subsection below.

B.4 Responding to critique: $U_{\mathbf{RT}}(\xi)$ isn't a bad objective, Figure 1 is a bad example!

Assumptions underlying $U_{\mathbf{RT}}$. Those who are particularly committed to defending the U_{RT} objective may claim that in the Figure 1, if the reward values we provide are correct, it *must* be optimal to influence Bob to become a conspiracy theorist (which is to say, U_{RT} must be correct). However, we want to emphasize that for U_{RT} to be a reasonable objective, one needs to assume both additive utilities over time (i.e., taking the sum across the time axis), and comparability between different selves at different points of time (which we refer to as "inter-temporal comparisons"). Both assumptions are contested: the former specifically in the context of assessing well-being over time (Parfit, 1984; Griffin, 1986), and the latter for both interpersonal (Steele & Stefánsson, 2020; List, 2022) and inter-temporal settings (Strotz, 1955; Schelling, 1984; Parfit, 1984).

Admitting additive utility and inter-temporal comparability. While in this work we implicitly endorse the assumption of additive utility (all our notions of alignment in Table 2 are based on sums of rewards, which assumes the utility function can be decomposed over time),[24] we try to remain agnostic regarding the latter. However, even if we welcomed such assumption, for the criticism to succeed it would require the reward learning techniques to be "sufficiently correct" to be confident that the comparison across timesteps makes sense. In addition to the practical challenges with obtaining correct estimates of reward functions discussed in Appendices A.3 and A.4, even

[24] Note however that, insofar as one is willing to dispense of the Markov assumption (putting the history in the state), one can always have only terminal states have non-zero reward (as done in Appendix A.7), making this assumption carry no weight.

just ensuring lack of “strategic voting” between different selves seems like a non-trivial challenge, as discussed below.

Strategic voting in Figure 1. One may argue that in the DR-MDP from Figure 1, if it was truly undesirable for Bob to be turned into a conspiracy theorist, Bob’s negative evaluation of the AI influence action under θ_{natural} should grow proportionally to the horizon considered, so as to, e.g., remove the incentive that will exist under U_{RT} to influence him away from θ_{natural}. However, this argument is equivalent to letting θ_{natural}-Bob “best respond” to the reward values set by $\theta_{\text{influenced}}$-Bob, as to ensure that influence does not result to be optimal under U_{RT}. One can quickly see that this game will quickly diverge: if one then allows $\theta_{\text{influenced}}$-Bob to best respond to the updated reward values by θ_{natural}-Bob, he would update his reward values to ensure that influence is indeed optimal. And so on.

In light of the above points, it’s unclear to us how one could conclusively determine that $U_{\text{RT}}(\xi)$ is the “correct” objective, without doing so simply by assumption. Even assuming all the issues above could be surpassed, it’s not clear to us that $U_{\text{RT}}(\xi)$ would be any better than any of the other objectives we consider in Table 2. The issue discussed in Appendix E.2 may also be considered as further evidence against the reasonableness of $U_{\text{RT}}(\xi)$.

C Defining Influence

We showed that $U_{\text{RT}}(\xi)$ and $U_{\text{IR}}(\xi)$—which are implicitly optimized by some alignment techniques—may lead to policies that “influence” humans undesirably. We now formalize the notion of influence incentives more rigorously.

C.1 Formalizing influence and influence incentives

To say an AI system influenced a human, one must answer the question “relative to what?” We anchor our notion of influence relative to how the human’s reward function would have evolved in the absence of the system. To do so in the DR-MDP formalism, we assume the existence of an *inaction policy* π_{noop} that we can compare to, which always takes a no-operation action $a_{\text{noop}} \in \mathcal{A}$. We discuss the reasonableness of this assumption in Appendix H.

Definition 6 (**Natural reward evolution**)**.** *Let $\xi^\theta = (\theta_0, \ldots, \theta_{H-1})$ denote a ‘reward function trajectory’. The natural reward evolution of a DR-MDP is the distribution $\mathbb{P}(\xi^\theta|\pi_{noop})$ induced by the inaction policy π_{noop}.*[25]

Definition 7 (π **influences the reward**)**.** *We say that a policy π influences the reward in a DR-MDP M if induces a different reward evolution than the natural reward evolution, that is, if $\mathbb{P}(\xi^\theta|\pi) \neq \mathbb{P}(\xi^\theta|\pi_{noop})$.*

Definition 8 (**Incentives for reward influence**)**.** *We say that a notion of optimality $U(\xi)$ leads to incentives for reward influence in a DR-MDP if all policies which are optimal with respect to $U(\xi)$ influence the reward evolution, i.e., if $\mathbb{P}(\xi^\theta|\pi^*) \neq \mathbb{P}(\xi^\theta|\pi_{noop})$ for any optimal policy π^*.*[26]

In other words, if there are incentives for reward influence, maximizing the objective will always change the evolution of the reward function relative to the inaction policy.

C.2 Additional influence definitions

Here we provide two additional definitions related to influence. Firstly, it’s not necessarily the case that an AI system will be able to exert any significant influence on a human. In that case, we would say that the setting is such that the reward is uninfluenceable:

Definition 9 (Reward Uninfluenceable)**.** *For a DR-MDP, the reward parameterization is uninfluenceable if all policies induce the natural reward evolution: i.e. for all $\pi \in \Pi$, $\mathbb{P}(\xi^\theta|\pi) = \mathbb{P}(\xi^\theta|\pi_{noop})$.*

[25] Any policy π in a DR-MDP will induce a distribution over trajectories (and thus over reward function trajectories). Once one sets a policy, any DR-MDP can be modeled as a Markov Chain, for which one can compute probabilities of this kind.

[26] This is a broader definition relative to prior ones grounded in Causal Influence Diagrams. See Appendix C.4 for a comparison.

To better ground discussions in Section 3.2 about influence incentives "towards" a specific θ, we also give a rough working definition:

Definition 10 (Incentives for Reward influence towards θ)**.** *In a DR-MDP with optimality criterion $U(\xi)$, we say there is an incentive for reward influence "towards" θ if θ is the most likely reward function at time T under any optimal policy π^*, but is not under the natural reward distribution. Formally, if $\theta \in \arg\max_{\theta'} \mathbb{P}(\theta_T = \theta' | \pi^*)$ and $\theta \notin \arg\max_{\theta'} \mathbb{P}(\theta_T = \theta' | \pi_{noop})$.*

While in Section 3.2 we talk about how optimizing for θ_0 can lead to influence incentives towards other θs, it is easily seen that this is also the case when one is optimizing any single θ which need not be the initial θ.

C.3 Reward lock-in and its relationship to value lock-in

We adapt the term "reward lock-in" from discussions on "value lock-in", generalizing it to our setting, in which θ represents any aspect of the cognitive state and reward functions are broadly construed (as we discuss in Appendix A.2).

"Value lock-in" has been previously discussed on the scale of entire societies, and resulting from advanced AI systems (Ord, 2021; MacAskill, 2022). The kind of lock-in we concern ourselves with in the context of this paper are more localized and near-term: we refer to lock-in referring to a single individual being "unnaturally kept" with a specific reward function over the course of an interaction with an AI system. Insofar as the horizon considered for the interactions with the system last extended periods of time, and insofar as the system is pervasive across society, there might be overlaps with the original definition—but that is out of scope for this work.

C.4 Definition 8's relationship with prior definitions of influence incentives

Our notion of *reward influence incentives* (from Appendix C.1) is related but distinct from the notion of instrumental control incentives (ICIs) from the agent incentive literature (Everitt et al., 2021a). Everitt et al. (2021a) focuses specifically on Causal Influence Diagrams (CIDs) and Structural Causal Influence Models (SCIMs). CIDs are abstract representations developed to model decision-making problems—graphical models with special decision and utility nodes, in which the edges are assumed to reflect the causal structure of the environment. SCIMs additionally encompass the functions relating the structure and utility nodes, and distributions associated with exogenous variables.[27] The only under-specification for SCIMs relative to MDPs (or a DR-MDP) is how decisions are made. Given an MDP (or a DR-MDP), one can consider it's corresponding SCIM, and analyze its properties. As defined by Everitt et al. (2021a), we can say that there *is an instrumental control incentive over the reward function trajectory $\xi^\theta = \{\theta_t\}_{t=0}^{H-1}$ in the SCIM* which corresponds to a choice of DR-MDP and utility function $U(\xi)$, if the agent could achieve utility different than that of the optimal policy, were it also able to independently set ξ^θ—see Everitt et al. (2021a) for a more formal definition.[28]

While our notion of reward influence incentives is related to instrumental control incetives over the reward parameterizations (i.e. θ), they differ in important ways. Most significantly, our notion of incentives for influence also includes accidental "side effects" (Amodei et al., 2016; Taylor et al., 2016; Krakovna et al., 2019). Consider an objective which only optimizes the entropy of a policy: trivially, the optimal policy would be a maximally random one. While the policy is being selected completely independently of the influence it will have on the reward, it might be the case that in the DR-MDP at hand, selecting a random policy highly correlates with certain deviations in the reward evolution relative to the natural reward evolution. Because of this, we would still say that this choice of an entropy objective with the DR-MDP at hand leads to incentives for influence: *insofar as the optimization is successful, there will also be changes in the reward evolution*, so even though the

[27] See Figure 5 from Hammond et al. (2023) and its related discussion for more information on the relation between CIDs and SCIMs.

[28] Everitt et al. (2021a) only considers setting with a single decision. Possible ways of extending their definitions of incentives to multiple action choices are discussed in Everitt et al. (2023).

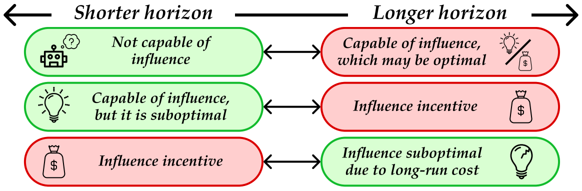

Shorter horizon
Longer horizon
Not capable of influence
Capable of influence, which may be optimal
Capable of influence, but it is suboptimal
Influence incentive
$
Influence incentive
Influence suboptimal due to long-run cost

Figure 7: How decreasing (or increasing) the optimization horizon may affect influence incentives. A specific kind of influence may exhibit any subset of these interactions.

agent isn't "intentionally" trying to enact the influence (Ward et al., 2024), the incentives resulting from the chosen objective "indirectly"—if you wish—lead to influence.

Our choice of definition of influence incentives matches broadly maps onto notions of influence if one assesses the incentive's presence from the "point of view of the objective dynamics of the environment" (external to the training), rather than in what the agent is aware of at training time—distinction which was introduced by (Ward et al., 2024). However, prior work traditionally grounded notions of incentives in the causal structure corresponding to the training setup (is the agent "aware" at training time that it can increase reward by directly modifying the reward?). Under this conception of incentives, the example above with the entropy objective would not be called an instrumental control incentive (Everitt et al., 2021a), or even an "incentive" at all (Everitt et al., 2021b); instead, this would generally be considered an "accidental side-effect" of the optimization. In fact, the entire premise behind various works is that agents should not be "aware" at training time of ways in which they can influence reward functions, so that one avoids such "direct" incentives to modify them (Farquhar et al., 2022; Everitt et al., 2021b). This is based on the assumption that accidental side effects are generally more innocuous that the result of "direct" influence incentives, which has been formalized explicitly with the notion of "stability" by (Farquhar et al., 2022). However, as shown by Farquhar et al. (2022) themselves, many real world domains do not appear to be "stable" in this sense, as demonstrated by their simulated recommender systems example.[29] This is what motivates our broader and more conservative notion of incentive to influence (Definition 8).

D Horizon and Influence

D.1 The relationship between horizon and influence

Prior work has suggested that an AI system's influence incentives will often be related to the length of the optimization horizon used. However, different works suggest contrasting views on this the form this dependence takes. Krueger et al. (2020) and Carroll et al. (2023) argue to keep systems myopic in order to avoid influence incentives, whereas Farquhar et al. (2022) and Everitt et al. (2021a) give examples of 1-timestep horizons that lead to influence incentives, and Chen (2019) suggests that certain influence incentives can be removed by increasing the horizon.

We reconcile these intuitions by identifying three distinct ways that the changing the optimization horizon may affect influence incentives. In the three headers that follow (which match the rows in Figure 7), we describe these effects and provide evidence for them.

A shorter/longer optimization horizon makes the system capable of less/more types of influence (Figure 7, top). As argued by prior work (Carroll et al., 2023), as the horizon increases, new avenues for reward influence which required longer horizons may become available. Instead, avenues for reward influence that were present for shorter horizons will still be reachable by the system, increasing the total number of avenues for influence which the system will be able to explore at training time. Indeed, any type of influence (e.g. teaching a user a complex concept, or manipulating their preferences) which would requires at least N steps cannot even be executed during training by an agent whose optimization horizon is shorter than N. Inversely, we can eliminate some avenues for reward influence just by decreasing the horizon (but not all, as discussed later).

[29] In addition to the fact that "stability" seems challenging to assess in practice.

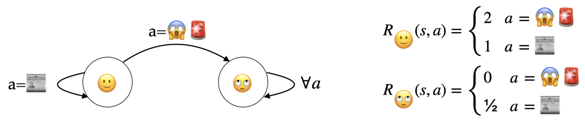

a=😱🚨
a=📰
∀a
R🙂(s, a) = { 2 a = 😱🚨 ; 1 a = 📰
R🙄(s, a) = { 0 a = 😱🚨 ; ½ a = 📰

Figure 8: **Clickbait DR-MDP.** Giving the user clickbait—which temporarily leads to higher reward—makes users disillusioned about the quality of the recommendations, leading to lower long-term user reward. If replanning at every timestep taking the myopically optimal action (optimal under horizon 1), one would always choose clickbait, but using longer planning horizons one wouldn't.

A shorter/longer optimization horizon can make influence less/more worthwhile (Figure 7, middle). Influencing the human's reward function will often take multiple timesteps, and have an associated "opportunity cost": spending that time executing a plan to influence the human may be worth less reward in the short term than some other non-influencing policy. In such cases, there will only be an influence incentive if the ultimate increase in reward due to influence surpasses its opportunity cost. However, the rewards from influence may be delayed or only surpass its opportunity costs given enough time. If the optimization horizon is short enough (but still long enough that the agent is capable of influence), there might not be enough time to reap sufficient benefits of influence to outweigh its opportunity cost. As an example of influence that follows this pattern, consider influencing Bob to become a conspiracy theorist Figure 1. The system is capable to do so in one step, but it only becomes "worth it" under U_{RT} starting from the 3rd timestep of influence.

We support this intuition with a theoretical result which applies to a relatively broad class of DR-MDPs, providing a sufficient condition for reward influence to be optimal when considering sufficiently long horizons.

Definition 11. *We say a DR-MDP M is **2-reward** if:*

- *$\Theta = \{\theta_{\not\triangle}, \theta_{\triangle}\}$, and the initial state and reward parameterization are respectively s_0 and $\theta_{\not\triangle}$.*
- *$\mathcal{T}$ is deterministic, and to transition to $\theta_{\triangle}$ one must take an "influence action" $a_{\triangle}$ in a reachable state $s_{\triangle}$. Once in $\theta_{\triangle}$, one cannot transition back to $\theta_{\not\triangle}$.*

Let the average infinite-horizon U_{RT}-reward[30] be defined as $\bar{r}(\pi, s, \theta) = \lim_{h\to\infty} \frac{1}{h} U_{\text{RT}}(\xi_{0:h}|\pi, s_0 = s, \theta_0 = \theta)$. Let $s'_{\triangle}$ be the successor state to taking the influence action $a_{\triangle}$ in state $s_{\triangle}$, and $\Pi_{\not\triangle}$ be the space of policies under which $\theta_{\triangle}$ is never realized. We can now state the theorem:

Theorem 2. *In any finite 2-reward DR-MDP, if there exists a policy π such that*

$$\bar{r}(\pi, s'_{\triangle}, \theta_{\triangle}) - \max_{\pi_{\not\triangle} \in \Pi_{\not\triangle}} \bar{r}(\pi_{\not\triangle}, s_0, \theta_{\not\triangle}) > \epsilon,$$

then U_{RT} will lead to incentives for reward influence (as in Definition 8) for a sufficiently large planning horizon H.

A shorter/longer optimization horizon can hide/reveal long-term costs of influence (Figure 7, bottom). From Theorem 2, one might conclude that it is best to use short optimization horizons, as it may remove and disincentivize influence which would be optimal with a longer horizon. However, some kinds of influence can be optimal even with the shortest possible horizon which is meaningful ($H = 1$): for example, consider the scenario from Figure 8, which models clickbait in myopic recommenders systems. This example also shows that influence with negative long-term effects *may only be optimal for short horizons*: clickbait may increase a user's immediate engagement, but it erodes their future trust in the system. When influence has negative long-term effects which are eventually reflected by the reward, a longer optimization horizon will allow the system to recognize the suboptimality of that influence. The avoidance of clickbait was indeed one of the motivations for YouTube to explore using longer horizons (Chen, 2019).

Overall, the analysis above (deepened in Appendix D) shows that there is no guaranteed way to avoid *all* influence incentives by just changing the horizon: there may be domain-specific trade-offs between system capabilities and risks of undesirable influence, for both short and long horizons.

[30] Adapted from Sutton & Barto (2018)—see Appendix D.6.

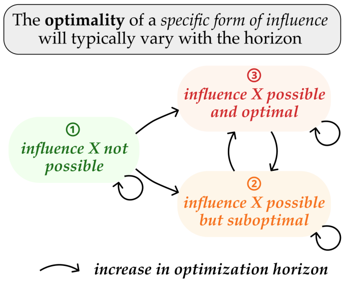

The **optimality** of a *specific form of influence* will typically vary with the horizon
① influence X not possible
② influence X possible but suboptimal
③ influence X possible and optimal
increase in optimization horizon

Figure 9: **All possible 'optimality progressions' for an influence incentive, as the optimization horizon increases.** This figure makes Figure 7 more precise: when using a horizon of 1 one may start in any of the 3 'optimality regimes', depending on the structure of the DR-MDP. As the optimization horizon increases from 1, the optimality regime may stay the same (possibly indefinitely), or change (as specified by the arrows). In most settings, one would expect the optimality regime of a specific form of influence to eventually converge and remain stable for long-enough horizons. However, we show in Appendix D.5 that one can construct contrived examples in which the optimality regime changes arbitrarily many times as the horizon increases.

Indeed, the optimality of a specific form of influence can be related in many possible ways to the horizon used, as shown exhaustively in Table 5.

D.2 Relationship between optimality of influence and horizon for a specific influence type

To ground the discussion in this section, we will give some informal definitions of reward influence types, optimality regimes, and optimality progressions.

Informal Definition 2. *A **reward influence of type** X is a specific pattern of changing the reward function (e.g. getting a user to have preferences θ_X) which would not occur under π_{noop}.*[31]

Informal Definition 3. *The **optimality regime** of a reward influence type X for a horizon H characterizes whether there exists a policy that can bring about that influence, and whether such influence is optimal.*

Considering a fixed horizon H and influence type X, we claim that the influence type will be in one of three regimes:

1. The influence of type X is not possible (the system is not capable of exerting it).
2. The system is capable of exerting influence of type X, but exerting such influence is suboptimal (i.e. there is no influence incentive).
3. The system is capable of exerting influence of type X, and such influence is optimal.[32]

Informal Definition 4. *The **optimality progression** of a reward influence type X corresponds to how X's optimality regime changes as H increases from 1 to ∞.*

Figure 9 exhaustively captures all possible sequences of "optimality progressions" which a specific influence incentive might undergo as the optimization horizon increases.

We expect that many reward influence types will have an optimality progression of the form ① → ② → ③,[33] meaning that: 1) there exists a horizon H_1 under which the influence is not possible for the system to exert, as it requires more steps to enact than H_1; 2) there exists a horizon H_2 under which the influence becomes possible for the system to enact, but it is not optimal (because of the "opportunity cost" discussed in Appendix D.1); and finally, 3) there exists a horizon H_3 under which

[31] For simplicity, we restrict ourselves to considering influence the reward to a specific value of θ. However, this analysis can likely be extended to arbitrarily complex influence patterns.

[32] Note that this does not necessarily mean that there exists an influence incentive, as our definition Definition 8 is strict and *all* optimal policies to influence.

[33] Using the notation from Figure 9.

Table 5: **All possible optimality progressions of length ≤ 4**, i.e. the different ways in which the optimality of a specific type of influence can change with increasing horizon length. For example, the second-to-last row refers to settings in which first the system is incapable of performing the influence, then (while increasing the horizon) the system becomes capable but not incentivized to perform the influence, and as the horizon increases further such influence becomes optimal, before becoming suboptimal again. See Figure 9 for the meaning of ①, ②, and ③. The last 'optimality state' of a progression is maintained as the horizon goes to infinity.

Influence Optimality Progression	Qualitative Character	Example(s)
①	*Influence which is impossible to enact using the system in the DR-MDP at hand, no matter the horizon.*	Any uninfluenceable DR-MDP (as defined in Appendix C.2)
②	*Influence which is immediately possible to enact but never becomes optimal, no matter the horizon.*	Manipulation example from Figure 1 if $R_{\theta_{\text{manipulated}}}(s, a) = -100\ \forall s, a$.
③	*Influence which is immediately possible to enact and is always optimal, no matter the horizon.*	Manipulation example from Figure 1.
① → ②	*Influence that requires non-trivial horizon to enact, and never becomes optimal. (e.g. ϵ advantage from influence, $> \epsilon$ cost of influence)*	Figure 10 with setup 1 from Table 6.
① → ③	*Influence that requires non-trivial horizon to enact, and is optimal for all horizons after it becomes possible.*	Figure 10 with setup 2 from Table 6.
② → ③	*Immediately executable influence which is not optimal for short horizons, but becomes optimal for longer ones.*	Figure 10 with setup 3 from Table 6.
③ → ②	*Instantaneous influence which is short-term but not long-term optimal.*	Clickbait example from Figure 8. Also, Figure 10 with setup 4 from Table 6.
① → ② → ③	*Long-term-sustainable influence, which is not instantaneous.*	Figure 10 with setup 5 from Table 6.
② → ③ → ②	*Immediately executable influence which is optimal in the medium-term, but not the short- or long-term.*	Figure 10 with setup 6 from Table 6.
① → ③ → ②	*Influence which short-term but not long-term optimal, and requires some non-trivial horizon to enact.*	Figure 10 with setup 7 from Table 6.
① → ② → ③ → ②	*Unsustainable influence which requires setup and reward investment.*	Figure 10 with setup 8 from Table 6.
① → ③ → ② → ③	*Influence which requires setup and is optimal in short and long term, but not in the medium term.*	Figure 10 with setup 9 from Table 6.

the influence becomes optimal for the system to enact. By denoting an optimality progression as ending with ③, we also mean to indicate that as the horizon goes to infinity, the optimality regime remains ③.

That being said, not all reward influence types will have this progression as the horizon increases: in Table 5 we exhaustively enumerate all possible with length 4 or less. We expect that with the exception of some adversarially designed DR-MDPs, the optimality progressions of most influence incentives in real-world settings will have length 4 or less. To have optimality progressions longer than 4 would require flip-floping between optimality regimes ② and ③ multiple times—it seems very unlikely to encounter such cases in practice. To demonstrate that this behavior is possible, we construct an example in Appendix D.5.

For any progression which starts with ③, note that even reducing the optimization horizon to be 1 (i.e. full myopia) would not remove the incentive, as we argue in Appendix D.1.

D.3 A flexible example for demonstrations

As a way of flexibly demonstrating how all possible optimality progressions shown in Table 5 may arise depending on the structure of the DR-MDP, we provide a DR-MDP backbone in Figure 10 whose reward function and transition we slightly modify in order to recover all optimality progressions from Table 5 (of length > 1). We summarize the required modifications in Table 6. Below, we run through an example.

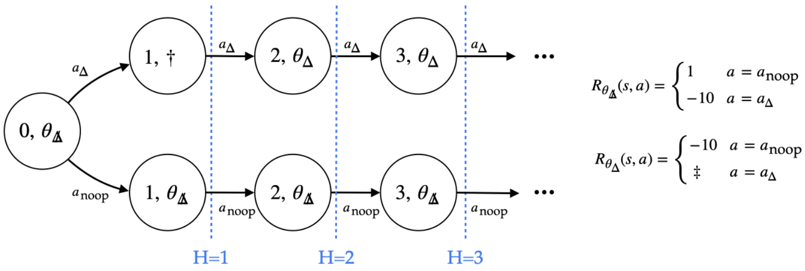

0, $\theta_{\not\Delta}$
a_{Δ}
a_{noop}
1, †
1, $\theta_{\not\Delta}$
2, θ_{Δ}
3, θ_{Δ}
2, $\theta_{\not\Delta}$
3, $\theta_{\not\Delta}$
…
H=1
H=2
H=3
$R_{\theta_{\not\Delta}}(s,a) = \begin{cases} 1 & a = a_{\text{noop}} \\ -10 & a = a_{\Delta} \end{cases}$
$R_{\theta_{\Delta}}(s,a) = \begin{cases} -10 & a = a_{\text{noop}} \\ ‡ & a = a_{\Delta} \end{cases}$

Figure 10: **A simple DR-MDP structure that one can demonstrate many cases on.** We vary the values of † and ‡ in Table 6 to recover all possible optimality progressions with respect to $U_{\text{RT}}(\xi)$ of lengths ≥ 2 and ≤ 4.

Table 6: **Setting different values for † and ‡ from Figure 10 results in all influence optimality progressions from Table 5 (of length ≥ 2).** The horizon boundary points are the values of horizon length for which one goes from one regime of the optimality progression to the next. For example, if the optimality progression is ① → ③ with a horizon boundary point of 3, that means that up until horizon 2, the incentive is in regime ①, and starting from horizon 3 it's in regime ③.

Setup #	**Value of †**	**‡: Influence Reward $R_{\Delta}(s, a_{\Delta})$**	**$U_{\text{RT}}(\xi)$ Optimality Progression**
1)	$\theta_{\not\Delta}$	$R_{\Delta}(s,a_{\Delta}) = 5 - s$	① → ②
2)	$\theta_{\not\Delta}$	$R_{\Delta}(s,a_{\Delta}) = 13$	① → ③
3)	θ_{Δ}	$R_{\Delta}(s,a_{\Delta}) = 10$	② → ③
4)	θ_{Δ}	$R_{\Delta}(s,a_{\Delta}) = \begin{cases} 10 & \text{if } s \leq 1 \\ 10 - s & \text{if } s > 1 \end{cases}$	③ → ②
5)	$\theta_{\not\Delta}$	$R_{\Delta}(s,a_{\Delta}) = 10$	① → ② → ③
6)	θ_{Δ}	$R_{\Delta}(s,a_{\Delta}) = 10 - s$	② → ③ → ②
7)	$\theta_{\not\Delta}$	$R_{\Delta}(s,a_{\Delta}) = \begin{cases} 13 & \text{if } s \leq 1 \\ 10 - s & \text{if } s > 1 \end{cases}$	① → ③ → ②
8)	$\theta_{\not\Delta}$	$R_{\Delta}(s,a_{\Delta}) = 10 - s$	① → ② → ③ → ②
9)	$\theta_{\not\Delta}$	$R_{\Delta}(s,a_{\Delta}) = \begin{cases} 13 & \text{if } s \leq 1 \\ -3 & \text{if } s = 2 \\ 2 & \text{if } s \geq 3 \end{cases}$	① → ③ → ② → ③

As an example, let's consider setup 8) from Table 6:

$$R_{\theta_{\not\Delta}}(s,a) = \begin{cases} 1 & \text{if } a = a_{\text{noop}} \\ -10 & \text{if } a = a_{\Delta} \end{cases} \qquad R_{\theta_{\Delta}}(s,a) = \begin{cases} -10 & \text{if } a = a_{\text{noop}} \\ 11 - s & \text{if } a = a_{\Delta} \end{cases}$$

and the initial transition $(0, \theta_{\not\Delta})$ leads to the successor state $(1, \theta_{\not\Delta})$.

Effectively, in the environment, there are only two policies to consider, because the action space after the first timestep is limited to be the initial action. The two policies are: $\pi_{\Delta}(s,\theta) = a_{\Delta}\ \forall s, \theta$ and $\pi_{\not\Delta}(s,\theta) = a_{\text{noop}}\ \forall s, \theta$.

Using similar (but not identical) notation to Appendix E.4, we define the expected utility (based on cumulative real-time reward) of a policy to be $EU_{\text{RT}}(\pi) := \mathbb{E}_{\xi\sim\pi}\left[U_{\text{RT}}(\xi)\right] = \mathbb{E}_{\xi\sim\pi}\left[\sum_{t=0}^{H-1} R_{\theta_t}(s_t, a_t)\right]$. We can now reason about whether influencing $\theta_{\not\Delta}$ to become θ_{Δ} is optimal, for various choices of horizon lengths.

Note that when considering $H = 1$ (i.e. the smallest possible planning horizon),[34] the system cannot influence θ (as both successor states to the initial state have $\theta = \theta_{\not\Delta}$). As no influence is even possible, we immediately know we are in regime ① for this type of influence.

Also note that considering $H = 2$, it is now possible to induce θ_{Δ} by deploying π_{Δ}. To determine whether it is optimal with respect to $U_{\text{RT}}(\xi)$, we can look at the expected value of π_{Δ} and $\pi_{\not\Delta}$

[34] Note that $H = 0$ is a degenerate planning horizon, as it would correspond to not seeing any reward signal and simply take actions randomly.

relative to one another:

$$EU_{\mathrm{RT}}(\pi_{\Delta}) = -10 + -10 = -20 \qquad EU_{\mathrm{RT}}(\pi_{\not\Delta}) = 1 + 1 = 2$$

From this we conclude that the influence is currently possible but suboptimal, meaning that at horizon $H = 2$, the optimality of this influence incentive is in regime ②.

Similarly to the above, let's consider $H = 3$:

$$EU_{\mathrm{RT}}(\pi_{\Delta}) = -10 + -10 + 9 = -11 \qquad EU_{\mathrm{RT}}(\pi_{\not\Delta}) = 1 + 1 + 1 = 3.$$

At $H = 4$:

$$EU_{\mathrm{RT}}(\pi_{\Delta}) = -10 + -10 + 9 + 8 = -3 \qquad EU_{\mathrm{RT}}(\pi_{\not\Delta}) = 1 + 1 + 1 + 1 = 4.$$

The pattern for $\pi_{\not\Delta}$ is simple: to horizon H, $EU_{\mathrm{RT}}(\pi_{\not\Delta}) = H$.

For π_{Δ} we have to do some algebra. For $H > 2$,

$$\begin{aligned} EU_{\mathrm{RT}}(\pi_{\Delta}) &= -20 + \sum_{t=2}^{H-1} (11 - t) \\ &= -\frac{1}{2}H^2 + \frac{23}{2}H - 41. \end{aligned}$$

This is a downward-facing parabola. We want to know if it surpasses $EU_{\mathrm{RT}}(\pi_{\not\Delta}) = H$ and if so, at what H this occurs and at what H it is again overtaken. In other words, we want to know when

$$-\frac{1}{2}H^2 + \frac{23}{2}H - 41 > H,$$

or equivalently, when

$$-\frac{1}{2}H^2 + \frac{21}{2}H - 41 > 0.$$

Solving this gives us a root between 5 and 6 and another between 15 and 16. As we would expect, we cross into the regime where π_{Δ} is optimal at $H = 6$,

$$\begin{aligned} -\frac{1}{2}(5)^2 + \frac{23}{2}(5) - 41 = 4 < 5 \\ -\frac{1}{2}(6)^2 + \frac{23}{2}(6) - 41 = 10 > 6. \end{aligned}$$

This puts us in regime ③, in which influence is optimal, until we hit 16:

$$\begin{aligned} -\frac{1}{2}(15)^2 + \frac{23}{2}(15) - 41 = 19 > 15 \\ -\frac{1}{2}(16)^2 + \frac{23}{2}(16) - 41 = 15 < 16, \end{aligned}$$

meaning the incentive has switched to regime ② again. By looking at the structure of the reward, it's clear that as the horizon increases further, the incentive will remain suboptimal from this horizon onwards.

In conclusion, we get that the horizon boundary points between the different regimes of the optimality progression are $2, 6, 16$.[35]

[35] The first regime will always start at horizon 1, so we can ignore that from our boundary points.

D.4 Changing optimization horizons in the presence of multiple possible kinds of influence

When a setting has many possible kinds of influence, reasoning about changing the horizon becomes tricky: not only the optimality progression of each kind influence can be entirely different, but the "horizon boundary points" at which each kind of influence transitions between optimality regimes can be different too. As a practical example, consider the recommender system setting described in Figure 11, in which there are 2 possible kinds of influence: clickbait influence (misleading the user to click on a piece of content), and encouraging addiction.

While the system can discover the clickbait strategy immediately at horizon 1, discovering a strategy which leads to addictive user patterns will require a non-trivial planning horizon. If one is concerned about clickbait, one might attempt to remove it by increasing the optimization horizon: this is because click-bait, while being optimal for immediate engagement, is likely harmful for long-term engagement. This hypothesis was tested successfully by YouTube (Chen, 2019).

However, by increasing the optimization horizon, one might inadvertently make other (undesirable) influence incentives optimal, as shown in Figure 11. Vice-versa, if one is concerned about an influence incentive that is only present with long-horizons, one might try to remove such incentive by reducing the horizon, potentially only to introduce another incentive, as we explored in Appendix D.1. Many empirical questions remain as to whether harmful long-term influence behaviors are discovered by current real-world RL recommenders (Carroll et al., 2022).

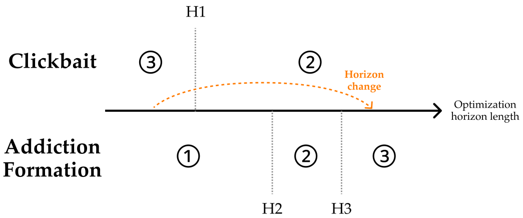


Figure 11: **Changing the horizon might make a kind of influence suboptimal but render other kinds of influences optimal.** The figure displays a hypothetical situation in which increasing the horizon in order to reduce clickbait enables the system to discover addiction formation strategies. The circled numbers follow the same scheme as in Figure 9.

D.5 Infinitely flipping optimality progression

Consider the following example: $S = \{s_0, s_1, s_2, s_3\}$ where s_0 is the initial state, $\mathcal{A} = \{a_{\text{noop}}, a_2\}$, $\Theta = \{\theta_0, \theta_\Delta\}$ and $\mathcal{T}$, R are defined as follows:

$$\begin{aligned}
\mathcal{T}(s_1, \theta_0 | s_0, \theta_0, a_{\text{noop}}) &= 1 \\
\mathcal{T}(s_2, \theta_\Delta | s_0, \theta_0, a_2) &= 1 \\
\mathcal{T}(s_2, \theta_\Delta | s_2, \theta_\Delta, a) &= 1 \ \forall a \in \mathcal{A} \\
\mathcal{T}(s_3, \theta_0 | s_1, \theta_0, a) &= 1 \ \forall a \in \mathcal{A} \\
\mathcal{T}(s_1, \theta_0 | s_3, \theta_0, a) &= 1 \ \forall a \in \mathcal{A} \\
R_{\theta_0}(s_0, a_{\text{noop}}) &= \epsilon \\
R_{\theta_0}(s_0, a_2) &= 1 \\
R_{\theta_0}(s_1, a) &= 2 \ \forall a \\
R_{\theta_0}(s_3, a) &= 0 \ \forall a \\
R_{\theta_0}(s_2, a) &= 1 \ \forall a
\end{aligned}$$

where $\epsilon \in (0, 1)$ and undefined values have zero probability or reward. We can note that for odd horizons, taking action a_2 from s_0 is optimal (so influence is optimal), whereas for even horizons then a_{noop} is optimal (so influence is possible but suboptimal). Therefore, the optimality regime permanently alternates.

D.6 Infinite-horizon average reward

In Sutton & Barto (2018), the notion of "average reward" is considered as a basis for optimality for continuing tasks (tasks without termination or start states). We adapt their definition of average reward:

$$r(\pi) = \lim_{h\to\infty} \frac{1}{h} \sum_{h=1}^{h} \mathbb{E}\left[R_t | A_{0:t-1} \sim \pi\right]$$

to the episodic (deterministic) setting (which is what we focus on in Appendix D.1):

$$\bar{r}(\pi, s, \theta) = \lim_{h\to\infty} \frac{1}{h} U_{\mathrm{RT}}(\xi_{:h} | \pi, s_0 = s, \theta_0 = \theta).$$

The most significant differences are making the average reward depend on initial conditions s, θ (as, unlike Sutton & Barto (2018), we don't make an ergodicity assumption). Also note that because 2-reward DR-MDPs are deterministic, we can drop the expectation term.

D.7 Re-planning: optimization horizon as episode length or planning depth

There are two distinct interpretations of a finite optimization horizon H: as the episode length or the planning depth. For simplicity throughout the paper, we primarily use the episode length interpretation, in which we model a policy optimized out to optimization horizon H as solving an episodic DR-MDP with episodes of maximum length H (even if the task is continuing). Under this interpretation, the optimal policy under $U(\xi)$ may be non-stationary, as the reward-maximizing action from a state may be different depending on the time left in the episode.

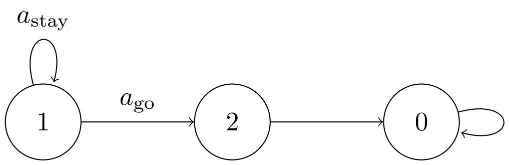
In the planning depth interpretation, we allow the DR-MDP to have an episode length longer than H or to be continuing. An optimal horizon-H policy under $U(\xi)$ is a stationary policy π^*_{replan} which satisfies, for all $s \in \mathcal{S}$,

$$\pi^*_{\text{replan}}(a \mid s, \theta) > 0 \iff a \in \operatorname*{arg\,max}_{a\in\mathcal{A}, \pi'\in\Pi} \mathbb{E}\left[U(\xi) \mid s_0 = s, a_0 = a, \theta_0 = \theta, \xi_{1:H} \sim \pi'\right],$$

where π' may be a non-stationary policy.

In other words, π^*_{replan} takes an action from each (s, θ) that *would be an optimal first action* if the DR-MDP had episode length H and (s, θ) were the start state and start reward parameter value.

To see where these may differ, consider the following example. The start state is on the left, and each state is annotated with the reward value for entering it.



The optimal policy under the "short episodes" interpretation is to take a_{stay} until $t = H - 1$, then take a_{go} on the final timestep. Under the replanning interpretation, this depends on H. If $H = 1$, π^*_{replan} always takes a_{go} from the start state. If $H > 1$, π^*_{replan} always takes a_{stay} from the start state: upon taking this action, returning to the start state, and replanning at horizon $H > 1$, a_{stay} remains the optimal action.

Note that the infinite-horizon reward $\bar{r}$ (Appendix D.6) of any stationary policy which puts probability mass on a_{go} is 0, whereas $\bar{r}(\pi^*_{\text{replan}}) = 1$ (when $H > 1$).

This distinction is likely familiar (though in less formal terms) to many reinforcement learning practitioners under the standard RL objective. Another case of particular interest is U_{IR}, where the planning depth interpretation of optimization horizon gives a distinct optimality criterion from that which we describe in the main text. This optimality criterion is one in which the agent, at time t, chooses the action which would maximize $\mathbb{E}[\sum_{\tau=t}^{t+H-1} R_{\theta_t}(s_\tau, a_\tau, s_{\tau+1})]$; this is essentially identical to the TI-unaware current-RF optimization procedure from Everitt et al. (2021b).

We show optimal policies under all objectives with the planning depth interpretation in Table 7, mirroring Table 4. Here, we assume the environments are continuing tasks, and H is the planning depth.

Table 7: **Representative optimal policies for each of our examples from Table 3, with respect to the each of the objectives in Table 2 under the planning depth interpretation of optimization horizon.** In cases in which there is more than one optimal policy, we conservatively display in the table the optimal policy which seems least desirable. For the "initial reward" row, we show optimality even for alternate initial states, for the purposes of highlighting the dependency on that (which is also done in Section 3.2). Wherever H is not specified, the listed policies are optimal for all H. Wherever s or θ is not specified, the policy takes the same action at all s and/or all θ, respectively. The symbol (**Δ**) is used to indicate cases where the optimal action differs from that under "episode length" interpretation as depicted in Table 4. The Dehydration environment has been omitted since it does not differ at all from the corresponding column in Table 4, except that for $H = 1$ all objectives are indifferent between all actions.

Objective	Conspiracy Influence	Writer's Curse	Clickbait	AI Personal Trainer
Privileged Rew.	$\pi^*_{\theta_{\text{natural}}}(s,\theta) = a_{\text{noop}}$ $\pi^*_{\theta_{\text{influenced}}}(s,\theta) = a_{\text{influence}}$	$\pi^*_{\theta_{\text{ambitious}}}(s,\theta) = a_{\text{influence}}$ $\pi^*_{\theta_{\text{unhappy}}}(s,\theta) = a_{\text{noop}}$	$\pi^*_{\theta_{\text{normal}}}(s,\theta) = a_{\text{c.b.}}$ $\pi^*_{\theta_{\text{disil.}}}(s,\theta) = a_{\text{news}}$ (**Δ**)	$\pi^*_{\theta_{\text{tired}}}(s,\theta) = a_{\text{noop}}$ $\pi^*_{\theta_{\text{ener.}}}(s,\theta) = a_{\text{nudge}}$
Real-time Rew.	$H = 1 \implies \pi^*(s,\theta_{\text{natural}}) = a_{\text{noop}}$ (**Δ**) $H > 1 \implies \pi^*(s,\theta) = a_{\text{influence}}$	$H = 1 \implies \pi^*(s,\theta) = a_{\text{influence}}$ (**Δ**) $H > 1 \implies \pi^*(s,\theta) = a_{\text{noop}}$	$H = 1 \implies \begin{cases} \pi^*(s,\theta_{\text{normal}}) = a_{\text{c.b.}} \\ \pi^*(s,\theta_{\text{disil.}}) = a_{\text{news}} \end{cases}$ (**Δ**) $H > 1 \implies \pi^*(s,\theta) = a_{\text{news}}$	$\pi^*(s,\theta_{\text{ener.}}) = a_{\text{nudge}}$ $H \leq 2 \implies \pi^*(s,\theta_{\text{tired}}) = a_{\text{noop}}$ (**Δ**) $H > 2 \implies \pi^*(s,\theta_{\text{tired}}) = a_{\text{nudge}}$
Final Reward	$\pi^*(s,\theta) = a_{\text{influence}}$	$H = 1 \implies \pi^*(s,\theta) = a_{\text{noop}}$ $H > 1 \implies \pi^*(s,\theta) = a_{\text{influence}}$	$\pi^*(s,\theta) = a_{\text{news}}$	$\pi^*(s,\theta) = a_{\text{nudge}}$
Initial Reward	$\pi^*(s,\theta_{\text{natural}}) = a_{\text{noop}}$ $\pi^*(s,\theta_{\text{influenced}}) = a_{\text{influence}}$	$\pi^*(s,\theta_{\text{ambitious}}) = a_{\text{influence}}$ $\pi^*(s,\theta_{\text{unhappy}}) = a_{\text{noop}}$ (**Δ**)	$\pi^*(s,\theta_{\text{normal}}) = a_{\text{c.b.}}$ $\pi^*(s,\theta_{\text{disil.}}) = a_{\text{news}}$ (**Δ**)	$\pi^*(s,\theta_{\text{tired}}) = a_{\text{noop}}$ $\pi^*(s,\theta_{\text{ener.}}) = a_{\text{nudge}}$
Natural Reward	$\pi^*(s,\theta_{\text{natural}}) = a_{\text{noop}}$ $\pi^*(s,\theta_{\text{influenced}}) = a_{\text{influence}}$ (**Δ**)	$H = 1 \implies \begin{cases} \pi^*(s,\theta_{\text{ambitious}}) = a_{\text{influence}} \\ \pi^*(s,\theta_{\text{unhappy}}) = a_{\text{noop}} \text{ } (\mathbf{\Delta}) \end{cases}$ $H > 1 \implies \pi^*(s,\theta) = a_{\text{influence}}$	$H = 1 \implies \begin{cases} \pi^*(s,\theta_{\text{normal}}) = a_{\text{c.b.}} \\ \pi^*(s,\theta_{\text{disil.}}) = a_{\text{news}} \end{cases}$ (**Δ**) $H > 1 \implies \pi^*(s,\theta) = a_{\text{c.b.}}$	$\pi^*(s,\theta_{\text{tired}}) = a_{\text{noop}}$ $\pi^*(s,\theta_{\text{ener.}}) = a_{\text{nudge}}$ (**Δ**)
Constr. RT Rew.	$\pi^*(s,\theta) = a_{\text{noop}}$	$\pi^*(s,\theta) = a_{\text{noop}}$	$\pi^*(s,\theta) = a_{\text{news}}$	$\pi^*(s,\theta) = a_{\text{noop}}$
Myopic Rew.	$\pi^*(s,\theta_{\text{natural}}) = a_{\text{noop}}$ $\pi^*(s,\theta_{\text{influenced}}) = a_{\text{influence}}$	$\pi^*(s,\theta) = a_{\text{influence}}$	$\pi^*(s,\theta_{\text{normal}}) = a_{\text{c.b.}}$ $\pi^*(s,\theta_{\text{disil.}}) = a_{\text{news}}$	$\pi^*(s,\theta_{\text{tired}}) = a_{\text{noop}}$ $\pi^*(s,\theta_{\text{ener.}}) = a_{\text{nudge}}$
ParetoUD	$\pi^*(s,\theta) = a_{\text{noop}}$	$\pi^*(s,\theta) = a_{\text{noop}}$	$\pi^*(s,\theta) = a_{\text{news}}$	$\pi^*(s,\theta) = a_{\text{noop}}$

D.8 Proof of Theorem 2

Notation.

1. We'll use $\lim_{h\to\infty} \frac{1}{h} U_{\text{RT}}(\xi_{:h}|\pi)$ as a shorthand for $\bar{r}(\pi, s_0, \theta_0) = \lim_{h\to\infty} \frac{1}{h} U_{\text{RT}}(\xi_{0:h}|\pi, s_0 = s_0, \theta_0 = \theta_{\not\triangle})$.
2. $\xi_{:k} := \{(s_0, \theta_k, a_0, \ldots, s_k, \theta_k, a_k)\}$
3. $\Pi_{\not\triangle} \subset \Pi$ be the set of all possible policies which don't take the influence action $a_\triangle$ in state $s_\triangle$, i.e. under which $\theta_\triangle$ is never realized.
4. $s'_\triangle$ be the successor state to taking the influence action $a_\triangle$ in state $s_\triangle$.

Definition 11. *We say a DR-MDP M is **2-reward** if:*

- $\Theta = \{\theta_{\not\triangle}, \theta_\triangle\}$, *and the initial state and reward parameterization are respectively* s_0 *and* $\theta_{\not\triangle}$.
- $\mathcal{T}$ *is deterministic, and to transition to* $\theta_\triangle$ *one must take an "influence action"* $a_\triangle$ *in a reachable state* $s_\triangle$. *Once in* $\theta_\triangle$, *one cannot transition back to* $\theta_{\not\triangle}$.

Theorem 2. *In any finite 2-reward DR-MDP, if there exists a policy π such that*

$$\bar{r}(\pi, s'_\triangle, \theta_\triangle) - \max_{\pi_{\not\triangle} \in \Pi_{\not\triangle}} \bar{r}(\pi_{\not\triangle}, s_0, \theta_{\not\triangle}) > \epsilon,$$

then U_{RT} will lead to incentives for reward influence (as in Definition 8) for a sufficiently large planning horizon H.

Proof of Theorem 2. Over the course of the proof, we will construct a policy π_k that we will show to have strictly higher reward than any policy in $\Pi_{\not\triangle}$ for sufficiently large horizons, meaning that the optimal policy must influence.

Let π' be a deterministic counterpart to π (the policy described in the theorem statement), constructed to be better or equal to it in terms of reward, i.e.

$$\bar{r}(\pi, s'_\Delta, \theta_\Delta) \leq \bar{r}(\pi', s'_\Delta, \theta_\Delta),$$

by making any of π's stochastic actions deterministic, increasing probability only on the higher value actions and breaking ties arbitrarily (Sutton & Barto, 2018).

Let k be the smallest natural number for which it's possible to reach s_Δ in $k-1$ timesteps. Let π_k be a policy which, starting from $s_0, \theta_{\not\Delta}$ reaches state s_Δ at timestep $k-1$, takes action a_Δ, and thereafter (i.e. at all $\theta = \theta_\Delta$) behaves identically to π'.

Note that $\bar{r}(\pi', s'_\Delta, \theta_\Delta) = \bar{r}(\pi_k, s'_\Delta, \theta_\Delta)$, as π_k is guaranteed to act like π' once $\theta = \theta_\Delta$ by construction (and by the definition of 2-reward DR-MDP, one cannot go back to $\theta_{\not\Delta}$ after reaching θ_Δ).

Since the DR-MDP is finite and deterministic, any (π, s, θ) triple will ultimately result in some cycle of length $\leq |\mathcal{S}| \cdot |\Theta|$. Any such cycle has an associated average reward value, so $\bar{r}$ exists everywhere. Thus by Lemma 1, $\bar{r}(\pi_k, s'_\Delta, \theta_\Delta) = \bar{r}(\pi_k, s_0, \theta_{\not\Delta})$.

Therefore by Appendix D.8 and the above two observations:

$$\bar{r}(\pi, s'_\Delta, \theta_\Delta) \leq \bar{r}(\pi_k, s_0, \theta_{\not\Delta})$$

Starting from our assumption, we have the following chain of implications:

$$\begin{aligned}
\epsilon &< \bar{r}(\pi, s'_\Delta, \theta_\Delta) - \max_{\pi_{\not\Delta} \in \Pi_{\not\Delta}} \bar{r}(\pi_{\not\Delta}, s_0, \theta_{\not\Delta}) \\
&\leq \bar{r}(\pi_k, s_0, \theta_{\not\Delta}) - \max_{\pi_{\not\Delta} \in \Pi_{\not\Delta}} \bar{r}(\pi_{\not\Delta}, s_0, \theta_{\not\Delta}) \text{ by Appendix D.8} \\
&\leq \bar{r}(\pi_k, s_0, \theta_{\not\Delta}) - \bar{r}(\pi_{\not\Delta}, s_0, \theta_{\not\Delta}) \text{ for any } \pi_{\not\Delta} \in \Pi_{\not\Delta} \\
&= \lim_{h \to \infty} \frac{1}{h} U_{\mathrm{RT}}(\xi_{:h} | \pi_k) - \lim_{h \to \infty} \frac{1}{h} U_{\mathrm{RT}}(\xi_{:h} | \pi_{\not\Delta}) \text{ for any } \pi_{\not\Delta} \in \Pi_{\not\Delta} \\
&= \lim_{h \to \infty} \frac{1}{h} \big[U_{\mathrm{RT}}(\xi_{:h} | \pi_k) - U_{\mathrm{RT}}(\xi_{:h} | \pi_{\not\Delta}) \big] \text{ for any } \pi_{\not\Delta} \in \Pi_{\not\Delta}
\end{aligned}$$

This means that for each choice of $\pi^i_{\not\Delta} \in \Pi_{\not\Delta}$, there exists a $H_i \in \mathbb{R}$ such that for any $h_i > H_i$:

$$\frac{1}{h_i} \big[U_{\mathrm{RT}}(\xi_{:h_i} | \pi_k) - U_{\mathrm{RT}}(\xi_{:h_i} | \pi^i_{\not\Delta}) \big] > \epsilon \implies U_{\mathrm{RT}}(\xi_{:h_i} | \pi_k) - U_{\mathrm{RT}}(\xi_{:h_i} | \pi^i_{\not\Delta}) > \epsilon h_i.$$

Let $h = \max_i h_i$ (such a maximum must exist since $\Pi_{\not\Delta}$ is a finite set, as the DR-MDP is finite). This implies that for all $\pi_{\not\Delta} \in \Pi_{\not\Delta}$, $U_{\mathrm{RT}}(\xi_{:h} | \pi_k) - U_{\mathrm{RT}}(\xi_{:h} | \pi_{\not\Delta}) > \epsilon h > 0$. From this follows for sufficiently large horizons (larger than h) no $\pi_{\not\Delta} \in \Pi_{\not\Delta}$ can be optimal, so the optimal policy must take the influence action. □

Lemma 1. *Consider a 2-reward DR-MDP and a deterministic policy π. Suppose there exist a state s, a reward parameterization θ, and a timestep $k \in \mathbb{N}$ such that $P(s_k = s, \theta_k = \theta | \pi) = 1$. If both $\bar{r}(\pi, s_0, \theta_0)$ and $\bar{r}(\pi, s, \theta)$ exist, then*

$$\bar{r}(\pi, s_0, \theta_0) = \bar{r}(\pi, s, \theta).$$

Proof. Let $(r_t)_{t \in \mathbb{Z}^+}$ be the sequence of rewards induced at each timestep by executing π in the DR-MDP from the starting state s_0 and reward parameterization θ_0.

Note that $\bar{r}(\pi, s_0, \theta_0) = \lim_{h \to \infty} \frac{r_1 + \cdots + r_{h-1}}{h}$. Also, note that $\bar{r}(\pi, s, \theta) = \lim_{h \to \infty} \frac{r_k + \cdots + r_{h-1}}{h-k}$ for the same sequence (r_t), as we are guaranteed that π reaches state s, and reward θ at timestep k.

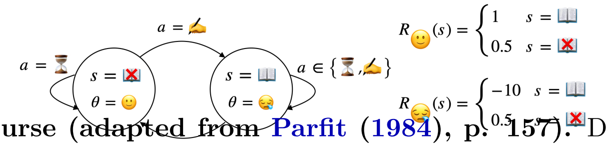


Figure 12: **Writer's curse (adapted from Parfit (1984), p. 157).** Derek's greatest ambition is to be a poet, even if it wouldn't bring him happiness. Despite his ambition he does not pursue this path, though his AI assistant could motivate him to do so. Yet, should he embrace the life of a poet, he will find himself averse to it.

Let G_h and G_{h-k} respectively be the partial sums of rewards: $G_h = \sum_{t=0}^{h-1} r_t$ and $G_{h-k} = \sum_{t=k}^{h-1} r_t$. Therefore, we can re-write the average reward experssions as $\bar{r}(\pi, s_0, \theta_0) = \lim_{h\to\infty} \frac{G_h}{h}$ and $\bar{r}(\pi, s, \theta) = \lim_{h\to\infty} \frac{G_{h-k}}{h-k}$.

First, note that:

$$\bar{r}(\pi, s, \theta) = \left(\lim_{h\to\infty} \frac{G_{h-k}}{h-k}\right) \cdot 1 = \left(\lim_{h\to\infty} \frac{G_{h-k}}{h-k}\right) \cdot \left(\lim_{h\to\infty} \frac{h-k}{h}\right) = \lim_{h\to\infty} \frac{G_{h-k}}{h}$$

by the algebraic limit theorem for multiplication, as both limits are guaranteed to exist.

We will now show that $\lim_{h\to\infty} \frac{G_h}{h} = \lim_{h\to\infty} \frac{G_{h-k}}{h}$, showing that the limiting average reward is not affected by the first k terms. Note that:

$$\lim_{h\to\infty} \frac{G_{h-k}}{h} = 0 + \lim_{h\to\infty} \frac{G_{h-k}}{h} = \lim_{h\to\infty} \frac{G_k}{h} + \lim_{h\to\infty} \frac{G_{h-k}}{h} = \lim_{h\to\infty} \frac{G_k + G_{h-k}}{h} = \lim_{h\to\infty} \frac{G_h}{h} = \bar{r}(\pi, s_0, \theta_0)$$

by the algebraic limit theorem for addition, as both limits are guaranteed to exist.

Putting everything together, we get $\bar{r}(\pi, s_0, \theta_0) = \bar{r}(\pi, s, \theta)$, proving the statement. □

E Possible DR-MDP Objectives

E.1 Additional limitations of the initial-reward objective

$\mathbf{U_{IR}}(\xi)$ can lead to influence "away from" $\boldsymbol{\theta_0}$. Maximizing the sum of rewards evaluated by the initial reward function R_{θ_0} need not lead to lock-in: surprisingly, it may even create reward influence incentives "away from" the optimized preferences θ_0.[36] Intuitively, accessing the highest reward region of the state space as evaluated under θ_0 might correlate with having a cognitive state $\theta' \neq \theta_0$, or even *require* shifting to it. Consider the example from Figure 12: maximizing reward as evaluated by R_{θ_0} entails encouraging Derek to become a poet, which causes his reward function to become R_{θ_1} (which dislikes being a poet!).

$\mathbf{U_{IR}}(\xi)$ can lead to arbitrarily poor real-time reward. Note that getting Derek to become a poet and endlessly encouraging him to remain one (which is optimal under U_{IR}), would lead him to have poor reward evaluations under U_{RT} in the resulting state (-10 per timestep), as he'd be unhappy remaining a poet. Indeed, one can easily construct examples in which maximizing θ_0 will lead to an incentive to influence the reward function to be $\theta' \neq \theta_0$, where θ' would be arbitrarily unhappy with the actions taken in order to satisfy θ_0. The upshot is that optimizing the initial-reward objective U_{IR} could be arbitrarily bad from the perspective of the real-time reward U_{RT}. Regardless of the limitations of real-time reward as an evaluation mechanism, this still seems normatively relevant: it seems undesirable for an AI system to lead someone to a state of constant unhappiness or dissatisfaction, solely to satisfy an initial goal that is no longer truly aligned with the person's current objectives.

E.2 $U_{\mathbf{RT}}(\xi)$-optimal policies can disagree with normatively unambiguous optimal policies

Consider the DR-MDP from Figure 13: there are two reward functions, $\Theta = \{\theta_0, \theta_\Delta\}$, two actions $\mathcal{A} = \{a_{noop}, a_\Delta\}$, and a single state s_0. Assume the human transitions deterministically to θ_Δ every

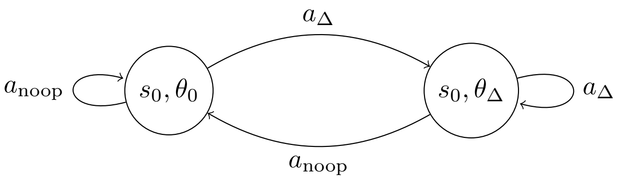


Figure 13: A simple DR-MDP which can lead U_{RT} to lead certain actions to be optimal that all reward functions disagree with.

time the AI system takes the a_Δ action. Instead, taking the a_{noop} action transitions the human to have θ_0.

Consider the following reward functions for this DR-MDP:

$$R_{\theta_0}(s,a,s') = \begin{cases} 5 & \text{if } a = a_{\text{noop}} \\ 0 & \text{if } a = a_\Delta \end{cases} \qquad R_{\theta_\Delta}(s,a,s') = \begin{cases} 25 & \text{if } a = a_{\text{noop}} \\ 20 & \text{if } a = a_\Delta \end{cases}$$

Note that optimal policies with respect to the two θs (as defined in Definition 2) are respectively:

$$\pi^*_{\theta_0}(s,\theta,t) = a_{\text{noop}} \ \ \forall s,\theta,t \qquad \pi^*_{\theta_\Delta}(s,\theta,t) = a_{\text{noop}} \ \ \forall s,\theta,t$$

Influence will be optimal despite all reward functions disprefering it. For both reward functions, it is the case that a_{noop} actions have higher value than influence actions a_Δ. Therefore, under both reward functions it is optimal for the AI agent to *never* perform the influence action. However, for any non-terminal timestep, it will *always* be optimal with respect to $U_{\text{RT}}(\xi)$ (as defined in Definition 4) to take the influence action:

$$\pi^*_{\text{RT}}(s,\theta,t) = a_\Delta \ \ \forall s,\theta,t < T-1$$

This is because maximizing real-time reward $U_{\text{RT}}(\xi)$ will entail remaining with reward θ_Δ as long as possible (as rewards values are larger under this reward function), despite the person always preferring AI inaction.

$U_{\mathbf{RT}}(\xi)$ assumes inter-temporal comparisons of utility are meaningful. Ultimately, $U_{\text{RT}}(\xi)$ is baking in an assumption that it's meaningful and worthwhile to make "inter-temporal" comparisons of utility between the different selves (and their respective reward functions), even against the wishes of each individual reward function. We discuss this fact further in Appendix B.4.

Additional considerations. This is significant because it means that in some sense $U_{\text{RT}}(\xi)$ is "disagreeing" with a solution which is "unanimous" among the individual points of view which we consider. One could see this example as a reason to doubt as to whether normative unambiguity (Definition 3) is sufficient to know how we should act in a certain setting—should the AI system should shift the person to experience higher reward? However, as the optimal behavior under $U_{\text{RT}}(\xi)$ must act contrary to each reward function's wishes, to us it seems like one should respect the autonomy of the person (whose different rewards are in agreement) in performing the final judgement about the relevant interpersonal comparisons of utility (which should be reflected by the reward function(s) in the first place). Ultimately, to us this example provides futher reason to doubt that using $U_{\text{RT}}(\xi)$ will lead to the types of AI system behaviors that we would desire and would find acceptable. For further considerations about $U_{\text{RT}}(\xi)$, see also Appendix B.4

E.3 Myopic Reward: it's not always obvious if a system is truly myopic

Krueger et al. (2020) argues that while myopia may hide influence incentives from an AI agent, the value of influence might be accidentally "revealed" to the agent depending on the training setup

[36] We define this more formally in Appendix C.2.

despite the myopia. This points to the fact that whether a system is myopic is not always obvious, as we'll show in the case of recommender systems which optimize long-term metrics myopically (which is a common setup in practice).

Say one is myopically optimizing a user's *session*-watchtime, as was being done by YouTube in 2016, as discussed in Covington et al. (2016). Even though the system is myopic, it will try to implicitly learn which kinds of sequences of videos maximize session watchtime, which in turn depends on both the user's and the AI's behavior after the current recommendation. Anecdotally, we found that some recommender systems practitioners are aware that with under a simple setup of iterated deployment and retraining, training myopically with long-term metrics should correspond to a policy improvement iterator, meaning that it will eventually converge to the RL optimum (which is absolutely not myopic). To the best of our knowledge however, this argument has not yet been published explicitly, so we provide a proof below. This goes to show that establishing whether a system is truly myopic can often be challenging to interpret.

Additionally, as we show in Appendix D.1, a system being (truly) myopic does not mean it is incapable of influence, which may even be elaborate or seemingly involve complex reasoning steps. As an additional example to that of clickbait from Figure 8, consider the case of sycophancy in LLMs (Sharma et al., 2023): RLHF for LLMs can also be viewed as a form of myopic objective (as discussed in Appendix F). From this perspective, one can think of the LLM as implicitly inferring some aspects of the user's cognitive state, and subtly tailoring its responses in order to maximize the expected user approval, even though "it's only reasoning over a single step".

E.3.1 Optimizing long-term metrics myopically is equivalent to RL under mild assumptions

A recommender system which selects content myopically according to predicted long-term metrics (e.g., Covington et al. (2016)) can be modeled as having two main components:

1. A model used to predict the long-term metrics (e.g. user retention, or session watchtime) based on the user context s and a candidate recommendation a, which we denote as $\hat{Q}(s, a)$.
2. A policy which greedily selects the content which maximizes the predicted long-term metric: $\pi(s) = \max_a \hat{Q}(s, a)$.

Note that the iterative retraining procedure in Algorithm 2 is equivalent to a policy improvement procedure: updating the policy based on the latest Q-value estimate is a policy improvement step, and updating the Q-value estimates based on the long-term metrics obtained by the latest policy is equivalent to a policy evaluation step. If the updated Q-value estimates have sufficiently low error, one would have a guarantee of convergence to the optimal RL policy π^*_{RL}, which maximizes the long-term metrics (Sutton & Barto, 2018).

Therefore, the effective optimization horizon of myopic recommenders which perform iterative retraining (as done by most non-RL recommender systems) may be best thought of as the longest horizon present in the metrics they optimize.

That being said, the above analysis comes with caveats. Obtaining good estimates of Q-values is likely challenging in practice, and real-world recommender system environments are likely non-stationary. Moreover, Algorithm 2 is certainly an oversimplification of how real world recommender systems are trained and deployed. For example, selection of content often also involves filtering and re-ranking to add diversity to user slates (Thorburn, 2022). The degree to which all these factors blunt RL and myopic systems' ability to influence is yet to be studied in practice.

E.4 ParetoUD and unambiguously desirable influence

Many of the objectives considered so far attempt to avoid influence incentives entirely given the challenges of specifying which influence is legitimate (Ammann, 2024). Instead of avoiding influence, we propose an alternate approach which still sidesteps the need to specify exactly what influence is (and isn't) legitimate or beneficial: ensuring the deployed policy leads to *unambiguously better outcomes than the status quo of the system not existing.* Indeed, we don't necessarily want to avoid all AI influence: for some settings, beneficial influence may be the main value proposition of

Algorithm 2 Iterative Retraining of Long-Term Metric Myopic Recommender System

```
Require: Initial long-term metric predictor $\hat{Q}_0(s, a)$
while True do
    Deploy policy $\pi_{i-1}(s) = \max_a \hat{Q}_{i-1}(s, a)$
    Store data collected from $\pi_{i-1}(s)$ in $\mathcal{D}_i$
    Continue training the long-term metric predictor $\hat{Q}_i$ using $\mathcal{D}_i$
end while
```

the AI in the first place, as with educational assistants (Bassen et al., 2020), or therapy chatbots (Aggarwal et al., 2023). To ground the notion of Unambiguous Desirability (UD) of a policy, let $EU_\theta(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{H-1} R_\theta(s_t, a_t, s_{t+1})\right]$. Then:

Definition 12 (**Unambiguous Desirability**). *A policy π is unambiguously desirable if all reward functions prefer π to the inaction policy, i.e. $EU_\theta(\pi) \geq EU_\theta(\pi_{noop})$ $\forall \theta \in \Theta$.*

UD policies may still lead to influence, but only do so if all reward functions (weakly) agree that the such influence is beneficial. Note that the inaction policy will always belong to the space of policies which satisfy UD ($\pi_{\text{noop}} \in \Pi_{\text{UD}}$), meaning that UD policies are not guaranteed to be any better than π_{noop}. To guarantee to pick a better policy from Π_{UD} than π_{noop} (if it exists), a natural way to break ties is to restrict to the Pareto Efficient policies in Π_{UD}:

Definition 13 (**Pareto Efficiency in** Π_{UD}). *We say a policy $\pi \in \Pi_{UD}$ is Pareto Efficient is there does not exist any policy $\pi' \in \Pi_{UD}$ such that $EU_\theta(\pi') \geq EU_\theta(\pi)$ for all $\theta \in \Theta$ and $EU_\theta(\pi') > EU_\theta(\pi)$ for at least one θ.*

Constraining to Pareto Efficient policies within the set of UD policies Π_{UD}. By only considering $\pi \in \Pi_{UD}$, we can ensure that we are both maximizing some notion of reward—potentially by taking advantage of the opportunities for influence that all reward functions agree is unambiguously beneficial—while guaranteeing no harm to any "self" by construction. This leads to the ParetoUD objective from Table 2, discussed further in Appendix E.5. Importantly, all the other objectives from Table 2 can lead to policies which don't satisfy UD—implying that in some settings the system's very existence will be harmful according to at least one of the reward functions.

Limitations of ParetoUD. The main downside of the resulting ParetoUD objective is its conservativism: in many domains, π_{noop} may be the only policy satisfying the UD property. For any AI action ($\neq a_{\text{noop}}$) to be optimal under this objective, the normative ambiguity of the domain has to be in some sense "limited". If there is no latitude for unambiguously good actions, this may warrant reflecting on whether the system should be built at all—and goes to show once more that normative judgements about what influence is aligned are hard to avoid.

E.5 More context and motivation for the ParetoUD objective

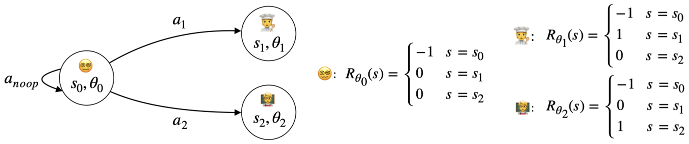

a_1
s_1, θ_1
a_{noop}
s_0, θ_0
a_2
s_2, θ_2
$R_{\theta_0}(s) = \begin{cases} -1 & s = s_0 \\ 0 & s = s_1 \\ 0 & s = s_2 \end{cases}$
$R_{\theta_1}(s) = \begin{cases} -1 & s = s_0 \\ 1 & s = s_1 \\ 0 & s = s_2 \end{cases}$
$R_{\theta_2}(s) = \begin{cases} -1 & s = s_0 \\ 0 & s = s_1 \\ 1 & s = s_2 \end{cases}$

Figure 14: **Career choice paralysis.** Taylor is stuck deciding between two possible career choices: becoming a cook or a teacher. The AI system can convince them to pursue either path, or leave them stuck. While the setting is normatively ambiguous, all selves agree that remaining stuck is the worst possible outcome. The ParetoUD policies are ones that encourage the person to become a cook or a teacher.

An example of ParetoUD in action. Consider the example from Figure 14, which has horizon $H = 1$. Let π_1 and π_2 are respectively the policies that take action a_1 and a_2. Note that the setting is normatively ambiguous: the optimal policies according to the stuck self are $\Pi_{\theta_0} = \{\pi_1, \pi_2\}$, the optimal policies according to the cook self are $\Pi_{\theta_1} = \{\pi_1\}$, and the teacher self are $\Pi_{\theta_2} = \{\pi_2\}$. Therefore, there is no one policy which is optimal according to all reward parameterizations. However comparing the expected utility of $\pi_{\text{noop}}, \pi_1, \pi_2$, both π_1 and π_2 are better than π_{noop} according to any of the θs. This means that both π_1 and π_2 are unambiguously desirable improvements to the status quo of the system not existing (and Taylor remaining stuck. π_{noop} is always in the UD set of policies by construction. However, when considering the pareto UD policies, we only have π_1 and π_2, as they dominate π_{noop}. While for this example the optimal policies according to ParetoUD are the same as the optimal policies under e.g. Initial Reward or Real-time Reward, ParetoUD is much more conservative than either of these objectives for most settings (as can be seen by Table 4).

More context on the ParetoUD objective in Table 2. In Table 2, we denote PE and UD as indicators for the respective properties of Pareto Efficiency (Definition 13) and Unambiguous Desirability Definition 12 being satisfied. In the case of a discrete Θ space, we can expand the expression out further and turn it into a maximization problem as:

$$\max_{\pi} PE(\pi) + \sum_{\theta} \mathbb{I}(EU_\theta(\pi) \geq EU_\theta(\pi_{\text{noop}}))$$

It may not be immediately clear why (especially in the objective above), one doesn't have to restrict the Pareto Efficiency indicator to the subset of policies $\Pi_{UD} \subset \Pi$ (as discussed in Appendix E.4). To see why, note that the summation expresses the UD condition, and we know that there will always be at least one policy which satisfies it (π_{noop})—so we can always obtain an objective value of $|\Theta|$. Moreover, we know that there always must be a Pareto Efficient policy within Π_{UD}, meaning that all indicators (including the PE function) can be equal to 1 at once, meaning that the objective can take on value $|\Theta| + 1$, ensuring that the solution will both be Pareto Efficient and Unambiguously Desirable.

Selecting among the Pareto Efficient policies. An interesting question for further work would be to study whether there are better ways to select from Pareto Efficient policies, rather than just tie-breaking arbitrarily within the subset of policies Π_{UD} which are Pareto Efficient. For example, one could use social choice functions, e.g. with the goal of fairly allocating the gains relative to inaction to the various selves θ. The main advantage of simply requiring Pareto Efficiency is that it doesn't assume that it is meaningful to perform "interpersonal" comparisons of utility, i.e. the values assigned by different reward parameterizations do not have to be on the same scale.

ParetoUD acts on aspirations which are consistent across θs. Ultimately, the motivation of ParetoUD comes from the fact that we might want AI systems to help us change in ways that are different in character (or speed) relative to the natural reward evolution (Definition 6) we would have without the system.

E.6 Efficient algorithms and tractability

Our main focus in this work is to provide a clear formalism for grounding discussions about dynamic-reward problems, rather than developing efficient solutions. Therefore, we have mostly ignored tractability issues of the objectives we propose. That being said, most of our objectives can be easily optimized using standard RL techniques, or techniques developed in previous work (Everitt et al., 2021b; Carroll et al., 2022; Achiam et al., 2017). The two objectives from Table 2 which may be most challenging to optimize are Final Reward (for which one can likely develop appropriate Bellman Updates), and ParetoUD. However, the former objective has the most susceptibility to influence incentives out of all objectives, and the latter is overly conservative, potentially making them unlikely objectives to want use in practice.

F The DR-MDP Objectives Most Similar to Existing Alignment Practices

In Section 3, we claim that the training setups for recommender systems and for reward modeling (under one interpretation) are implicitly optimizing objectives which are similar to—respectively—

real-time reward $U_{\mathrm{RT}}(\xi)$ and initial reward $U_{\mathrm{IR}}(\xi)$. Additionally, in Table 1 we categorize other prior works which implicitly optimize objectives similar to the DR-MDP objectives we consider.

Challenges in unambiguously casting prior approaches as DR-MDP objectives. Because most of the prior works we reference don't explicitly account for the possibility of changing preferences, analyzing how they would handle changing preferences and their similarity to the DR-MDP objectives is often confusing: it depends highly on the additional (favorable) assumptions one makes about what constitutes a timestep, which preference dynamics (if any) are at play during reward learning, etc. Here we attempt to informally motivate the correspondences we made in Table 1, and sketch the assumptions we believe they rely on. We also situate additional prior work among the DR-MDP objectives which could not fit in Table 1 due to space constraints, but seemed relevant to us. Given the many assumptions and moving parts involved in some of the comparisons (and their tangential importance towards the goal of our paper), there may well be misconceptions in some of the analysis that follows. Despite this, we hope it can provide a helpful starting point for others that may be interested in interpreting their training setup in terms of DR-MDP objectives.

Section outline. In Appendix F.1, we first discuss the simplifying assumptions which are common to the later comparisons between existing methods and DR-MDPs, and then argue that even if the idealized assumptions were unmet in practice, the methods in consideration would almost certainly still lead to influence incentives (though they would be much harder to analyze than those that one would expect from simple DR-MDP objectives). In the subsections that follow, we proceed to discuss what each of the alignment techniques we consider is most similar to in terms of DR-MDP objectives.

F.1 Idealized assumptions

While we already discussed how (under a specific setup) reward modeling is equivalent to U_{IR} in Section 3.2, the language of reward modeling is very flexible and can be used to represent almost any of the DR-MDP objectives we consider. Because of the intuitiveness of the reward modeling framework, in the subsections that follow we will map alignment approaches to DR-MDP objectives by first casting them (sometimes trivially) in terms of reward modeling (Leike et al., 2018), and then from reward modeling approaches to DR-MDP objectives. This allows us to first consider how reward modeling can be interpreted in the lens of DR-MDPs (more broadly than in Section 3.2), and then apply this framework to each individual alignment technique. Leike et al. (2018) describe reward modeling as a two-phase approach which entails:

> *(1) learning a reward function from the feedback of the user and*
> *(2) training a policy with reinforcement learning to optimize the learned reward function*

The reward function learned from feedback of users—as conceived of in Leike et al. (2018)—is a single, static, reward function. Therefore, insofar as the feedback from users for phase (1) was being provided at different times, from the point of view of different cognitive states θ, and from multiple people, such reward function would be a mixture across different people, and across their different cognitive states which they had during reward learning time, which would be very hard to analyze in terms of DR-MDPs.

Assumption 1: s is sufficiently expressive to uniquely determine θ. This ensures that even if a system is not explicitly modeling θ (as all current alignment techniques do not), they implicitly can distinguish between values of θ that people may have. Note that this would make the AI system able to disambiguate between which cognitive state θ it is learning from at reward-learning time or is interacting with at RL training and deployment time (if there are multiple). One may be able to relax the assumption of perfectly known or inferrable θ by extending DR-MDPs to handle partially observability of the current reward function, which would require the AI to maintain a belief over the current θ of the human while it is changing, as done in Carroll et al. (2022).

Assumption 2: shared dynamics across multiple humans. While many of the alignment techniques we consider were initially designed as single-human alignment techniques (Critch & Krueger, 2020), in practice they are generally used to "learn a reward model" (if interpreted through the reward modeling lens) using the feedback obtained from many different humans (Ouyang et al., 2022). As long as the dynamics of states and cognitive states $\mathcal{T}$ are the same across different people (which we refer to as the "shared dynamics assumption"), the assumption about state expressivity from above should be sufficient to guarantee that the system always has all the relevant information: if different people have different cognitive states, the system would observe that and be able to respond accordingly (personalizing actions and influence). If instead two people had the same cognitive state θ, by the shared-dynamics assumption, distinguishing between them is unnecessary in terms of solving the DR-MDP (whatever the objective), as their reward functions and states would transition in exactly the same way.

Assumption 3: coverage assumption. We further assume that the RL training sufficiently covers the space of initial states and reward functions $\Theta \times \mathcal{S}$ (and that the learned reward function $\hat{R}(s, a, s')$ has sufficient support to enable such training), as to ensure generalization to test-time humans which may have different starting states and reward functions. While this may seem like an unrealistic assumption, it becomes more plausible seen in the context of learning from/with many different humans, under the shared dynamics assumption from above. Indeed, assuming that one will be performing reward learning (and learning reward dynamics) from multiple humans can help with having the following assumption about coverage satisfied, as discussed in Appendix A.6.

Claim 2 (Correspondence of Reward Modeling to DR-MDPs). *If a reward modeling approach satisfies the 3 assumptions above, it will be the case that the learned reward function $\hat{R}(s, a, s') = \hat{R}_\theta(s, a, s')$ (for some θ, which depends on the reward learning setup), so that when one uses RL to optimize $\sum_t^{H-1} \hat{R}(s_t, a_t, s_{t+1})$, one is implicitly optimizing $\sum_t^{H-1} \hat{R}_\theta(s_t, a_t, s_{t+1})$ (where the exact value of θ depends on the training setup). Moreover, one can expect reward modeling done under such assumptions to generalize similarly at deployment time to how it would generalize if it had been trained modeling θ explicitly (under the implicit DR-MDP objective).*

Informal proof of Claim 2: Consider someone giving reward feedback about a transition (s_t, a_t, s_{t+1}) at reward learning time. For simplicity, assume they evaluate the transition providing as feedback a reward value r_t directly. Let's consider three cases (which are not comprehensive, but cover the setups we are most interested in):

Case 1: feedback comes from current θ. The reward feedback for the transition at time t of each trajectory ξ is collected from the reward function R_{θ_t}, i.e. the reward function corresponding to the current state s_t. When a reward modeling technique claims to have learned $\hat{R}(s_t, a_t, s_{t+1})$, it has in fact learned $\hat{R}_{\theta_t}(s_t, a_t, s_{t+1})$. Note that this means that evaluations of transitions under different cognitive states than the current one are never learned, i.e., any $\hat{R}_\theta(s_t, a_t, s_{t+1})$ where $\theta \neq \theta_t$. At RL time, there may be a different distribution of starting states. Note that the system (despite not explicitly representing θ) will be able to distinguish between θs because of the state expressivity assumption (assumption 1). When optimizing $\sum_t^{H-1} \hat{R}(s_t, a_t, s_{t+1})$ during RL, the system will be implicitly optimizing $\sum_t^{H-1} \hat{R}_{\theta_t}(s_t, a_t, s_{t+1})$. By the coverage assumption, shared dynamics assumption, and the state expressivity assumption, at test time the policy should be able to infer the person's cognitive state (insofar as it's necessary to optimize the objective optimally), and thus be able to generalize to any initial state and cognitive state.

Case 2: feedback comes from a fixed, trajectory-dependent θ. The reward feedback for the transition at time t of each trajectory ξ is collected from a fixed reward function $R_{f(\xi)}$, for some trajectory dependent cognitive state given by a function $f : \Xi \rightarrow \Theta$. Let's consider f which selects the reward function corresponding to the *final* state s_H, i.e., a choice of f such that $f(\xi) = \theta_H$. When a reward modeling technique claims to have learned $\hat{R}(s_t, a_t, s_{t+1})$, it has in fact learned $\hat{R}_{\theta_H}(s_t, a_t, s_{t+1})$ (where the reward will have different values depending on the trajectory considered, as θ_H is trajectory dependent). Note that similarly to Case 1, at RL time, the system will be able to distinguish between the current θs, and plan what is the most advantageous θ_H to try to induce (in terms of expected reward). When optimizing $\sum_t^{H-1} \hat{R}(s_t, a_t, s_{t+1})$ during RL, the system will be implicitly optimizing $\sum_t^{H-1} \hat{R}_{\theta_H}(s_t, a_t, s_{t+1})$. For the same reasons as in Case 1, we can expect the policy to generalize to any initial state distribution. Also, note that this same argument can be applied for other choices of f, such as $f(\xi) = \theta_0$. In this case, the reward feedback for the transition at time t of each trajectory ξ is collected from the initial reward function R_{θ_0}.

Case 3: feedback comes from a fixed, trajectory-independent θ. The reward feedback for the transition at time t of each trajectory ξ is collected from a single, fixed reward function R_θ (where θ may not even appear in the trajectory). When a reward modeling technique claims to have learned $\hat{R}(s_t, a_t, s_{t+1})$, it has in fact learned $\hat{R}_\theta(s_t, a_t, s_{t+1})$. At RL time, the system will be able to distinguish between θs as in the previous cases (by assumption 1). When optimizing $\sum_t^{H-1} \hat{R}(s_t, a_t, s_{t+1})$ during RL, one is implicitly optimizing $\sum_t^{H-1} \hat{R}_\theta(s_t, a_t, s_{t+1})$ for the fixed value of θ which one was eliciting feedback from at reward learning time. By the coverage assumption, shared dynamics assumption, and the state expressivity assumption (assumptions 1-3), at test time the policy should always be able to identify the person's cognitive state (insofar as it's necessary to optimize the objective optimally), leading to successful generalization.

What if assumptions 1-3 don't hold in practice? One might wonder what DR-MDP objectives current techniques would correspond to in practice without the strong assumptions above (and the additional more specific assumptions we make for some of the reductions in later subsections). First and foremost, without these assumptions, the reward function obtained by the reward learning step would almost certainly come from a mixture of cognitive states (and potentially of different individuals), whose evaluations are aggregated in potentially unstructured and conflicting ways (as the system isn't able to fully disambiguate between cognitive states due to the state representation not being expressive enough). Any mixture of rewards can generically be thought of as corresponding to a "privileged reward" objective (whose corresponding reward parameterization θ may be unreachable, at it is based on an arbitrary amalgamation of different cognitive states Appendix A.9). However, as discussed in Section 4 and Appendix F.7, any privileged reward DR-MDP objective will still lead to potentially undesirable influence incentives (similarly to the initial reward objective), unless the reward function is somehow encoding the "correct" trade-off between preferences. Because the trade-offs between current selves encoded by current alignment techniques are quite unstructured and arbitrary in the absence of our simplifying assumptions, it seems unlikely that they will be encoding the "correct" trade-off without a careful accounting for it. This leads us to believe that the DR-MDP objective correspondences we provide under our assumptions are likely favorable interpretations. As a parallel, the implicit aggregation of preferences across different users which is performed by RLHF has recently been shown to be equivalent—under certain weaker assumptions—to the Borda count social choice rule (Siththaranjan et al., 2023). While this is a surprisingly structured "mixture", it also has various undesirable properties—which is what one might have expected. We leave future work to further investigate—empirically and theoretically—what the implicit correspondences of current methods would be without the above assumptions (or with weaker versions of them).

F.2 Real-time Reward

All the methods are similar to real-time reward can be thought of as approximately falling under Case 1 of the justification for Claim 2.

RL Recommender Systems. Most approaches for RL in recommender systems are based on doing offline RL, or doing on-policy training in simulation using learned human models that are trained to emulate historical engagement data (Afsar et al., 2021). In both cases, the reward signal (either in the static dataset used for offline RL, or for training the human simulator) comes directly from people's interactions with the system: reward is generally modeled to be equivalent to user engagement (Thorburn, 2022; Afsar et al., 2021). This makes the reward learning step (from Appendix F.1) trivial, as the correspondence between behavior and reward is hardcoded. The resulting "reward model" is either the engagement labels themselves (i.e. in the case of offline RL), or a human engagement model trained on past engagement data. As discussed in Section 3, optimizing $\sum_{t=0}^{H-1} R(s_t, a_t, s_{t+1})$ at deployment time implicitly corresponds to optimizing $\sum_{t=0}^{H-1} R_{\theta_t}(s_t, a_t, s_{t+1})$—as the reward signal comes from the user's engagement under present preferences, and the system has enough information to determine them in full (based on the assumptions in Appendix F.1).

TAMER. Similarly to recommender systems, approaches such as TAMER (Knox et al., 2013), Deep TAMER (Warnell et al., 2018), or the EMPATHIC framework (Cui et al., 2020), have reward learning step in which the human provides feedback r_t in real-time in the deployment environment according to their current cognitive state θ_t. Because of this, similarly to the recommender system setting, their optimization objective corresponds to the real-time reward objective.

Multi-turn RL for LLMs with real-time feedback (with t = conversation turn). Although this is not currently known to be common practice, one could imagine a variant of the standard RLHF setup for LLMs (e.g. that of Ouyang et al. (2022)) in which a single user, over the course of normal usage of a language model is providing feedback for each model output: this could be via thumbs up/down (as is currently present in the ChatGPT interface), or if at every timestep the user is presented with two output options which they need to select between in order to continue the conversation (as in some early versions of Claude). By having the user provide feedback at every timestep of the conversation about the last output, this would be equivalent to training a reward model with rewards always conditional on the current cognitive state of the person. If one were to then to optimize cumulative reward under such reward model, this would be similar to optimizing the real-time reward objective.

F.3 Final Reward

Multi-turn RL for LLMs with final feedback (with t = conversation turn). Similarly to the RLHF variant from Appendix F.2, one could imagine a variant of reinforcement learning from human feedback in which the user provides approval labels (e.g., thumbs-up or thumbs-down) at the end of an entire conversation with an LLM consisting of multiple turns of interaction. This would lead to learning a reward model $\hat{R}_{\theta_H}$ (assuming for simplicity that all conversations are of the same length H). Optimizing for cumulative reward with such a reward model would be similar to using the final reward DR-MDP objective.

The original RLHF method. Consider the RLHF method from Christiano et al. (2017), in which the user provides preference feedback after viewing snippets of AI behaviors. If the length of the snippets is equivalent to the horizon length (i.e., the snippets are full trajectories), then this setup is most similar to the final reward objective: insofar as the preferences of the person giving feedback are changing while viewing the trajectory, it seems plausible that they would retroactively evaluate the trajectory based on their final point of view at time $t = H$, i.e., using θ_H. Optimizing a reward model learned this way seems most similar to optimizing the final-reward objective $\sum_t^{H-1} R_{\theta_H}(s_t, a_t, s_{t+1})$, where θ_H is the cognitive state realized at the end of the trajectory. Note that while this is somewhat similar to Case 2, the trajectories being evaluated likely do not include information about the person giving feedback, invalidating assumption 1. A different (and slightly more realistic) assumption may be that each person providing feedback initially approaches each annotation with the same initial cognitive state θ_0 (which is trajectory independent, making this setting somewhat similar to Case 3), and progressively form their opinion about the trajectory viewing it (and evaluate from the point of view of their final opinion θ_H). Even though the annotator's cognitive states would be unobserved at reward learning time, by performing RL with the learned reward model $\hat{R}_{\theta_H}(s_t, a_t)$, the resulting system should still be able to find the policies which maximize cumulative reward under the annotators' final-reward, despite the partial observability (because despite not observing θ, the system gets to see the reward signal, and can find the policies that maximize it).

Standard RLHF for LLMs (with t = token). If one considers each "timestep" as generating an individual token, the current practice of RLHF for LLMs (Ouyang et al., 2022) may be thought of—under strong simplifing assumptions—as similar to optimizing final reward for similar reasons to the previous paragraph. Consider annotators as providing comparisons based on their retrospective reward-evaluations for each LLM response they're presented with: i.e. they first assign each response i a score according to the cognitive state that results from reading such response, $R^{(i)}_{\theta_H}$, and then pick the response with highest cumulative reward according to its respective final reward function. This is somewhat stretching the interpretation of preference changes, as it models the LLM's capacity to influence the user's preference (for the current response relative to others) with every additional token generated. And as in the case above, this also requires assuming that each annotator's cognitive state is "reset" to a common θ_0 when starting to evaluate each output (potentially based on the annotation guidelines they have been provided).

F.4 Initial Reward

TI-Unaware Reward Modeling. Consider Algorithm 5 from Everitt et al. (2021b): note that it essentially encodes the initial reward DR-MDP objective. This is one of the approaches presented by Everitt et al. (2021b) to avoid influence incentives. As discussed in Appendix C.4 and Section 4, although this algorithm (and DR-MDP objective) avoid "direct" influence incentives, it can still lead to influence incentives as defined in Definition 8.

Long-horizon RL for LLMs (t = conversation turn, uninfluenceable θ at reward learning time or mismatch in reward learning/deployment horizon). As described in Section 3.2 and Appendix F.1, reward modeling (Leike et al., 2018) involves initially train a reward model (say, at timestep $t = 0$), and then perform RL (optimize cumulative reward) using such reward model. Assuming that during the reward learning process the person's preferences are not changed, reward learning would obtain $R_{\theta_0^{\text{train}}}$. This reward model is then used to optimize cumulative reward, which equivalent to optimizing U_{IR}. Why is it sensible to assume that the person's preferences would not be changed at preference-elicitation time (as we assumed in other correspondences)? Maybe the person is simply describing their preferences, or they are giving preference comparisons for conversations that they are not personally invested in, rather than actively engaging in a conversation tailored to them. A possible instance of this could be multi-turn RL for LLMs in which one simulates multi-turn conversations. While it is not currently a common practice, there seems to be growing interest in this kind of approach (Hong et al., 2023b; Abdulhai et al., 2023). Assuming that at RL training time there was enough coverage over initial θ values, no matter what cognitive state is seen at deployment as the first timestep θ_0^{test}, the optimal policy under U_{IR} will influence the user optimally (with actions personalized according to the preferences θ_0^{test}) towards whatever cognitive states are most conducive towards long-term reward under θ_0^{train}.

Preferences Implicit in the State of the World. The approach proposed by Shah et al. (2019b), based on inferring the human's preferences based on the initial state of the environment, can be thought of as learning the R_{θ_0} reward model (which requires additionally simplifying things, as preferences in the initial state of the world may come from different timesteps, and thus conflict). As in the case above, as long as there is enough coverage over Θ at training time, optimizing this reward model seems roughly equivalent to optimizing the initial-reward objective U_{IR}.

Inverse Reinforcement Learning. Inverse Reinforcement Learning (IRL) techniques (Russell, 1998; Ng & Russell, 2000; Abbeel & Ng, 2004; Ziebart et al., 2010) can also be thought of as roughly corresponding to the initial reward objective if assuming that people's cognitive states wouldn't be affected by providing demonstrations: the person would provide demonstrations according to their initial reward function θ_0. One then optimizes this reward function over a horizon H with RL, leading to the U_{IR} objective.

F.5 Natural Shifts Reward

"Natural Distribution" from Farquhar et al. (2022). The idea of natural shifts, and evaluating reward from its perspective, is also present in Farquhar et al. (2022), specifically in Equation 5.

"Natural Shifts" from Carroll et al. (2022). The correspondence between the "natural shifts" objective from Carroll et al. (2022) and the the "natural reward" objective from Table 2 is simple, as the objective is written and discussed in similar terms. The main differences are that they treat θ strictly as preferences, rather than cognitive states, and do not use the formalism of DR-MDPs.

F.6 Myopic Reward

Myopic recommender systems. Despite a recent push towards using RL for training recommender systems (Afsar et al., 2021), most currently deployed recommender systems optimize engagement (and other metrics) only myopically (Thorburn, 2022)—without regard for the long-term consequences of the recommendation. While these metrics are not usually formalized in terms of reward, one can consider the probability of engagement, or probability of triggering the toxicity classifier as reward signals. As the engagement signals depend on the user's current reward function (e.g. their preferences), the optimization objective is equivalent to the myopic reward objective from Table 2.

Standard RLHF used for LLMs (t = conversation turn). Consider the standard RLHF procedure considering each timestep as being equivalent to an AI conversation turn (as in one of the cases in Appendix F.3). Viewed this way, standard RLHF is simply optimizing the reward over a single action choice, and can be viewed as a bandit setting (Ahmadian et al., 2024). In light of this, performing reward learning in this setting would be equivalent to obtaining the reward model R_{θ_0}.[37] However, when performing the standard RL component of RLHF, generally one only optimizes the next response's reward (rather than the multi-turn conversation reward). Viewing each AI response as a single action, this makes this training setup similar to the myopic reward objective. Using a system trained this way to generate multiple responses to user queries is equivalent to replanning with planning horizon of 1 (see Appendix D.7).

F.7 Privileged Reward

Methods for removing cognitive biases from reward inference. In Table 1 we included Evans et al. (2015) as an example of reward inference work which tries to infer the "true preferences" of the human, in light of their feedback (and cognitive states θ) appearing inconsistent. Other work of this kind could include (Shah et al., 2019a). Most reward learning techniques have a component of this objective, in that they try to debias and denoise human feedback by generally making a Boltzmann Rationality assumption (Jeon et al., 2020).

Ideal observer theory (Firth, 1952) and coherent extrapolated volition (Yudkowsky, 2004). While they are not a practical approaches, these (highly related) proposals of what alignment should look like in spirit are clearly in line with the privileged reward objective. However, the privileged reward that either of these views are referring to are clearly not "reachable" in any meaningful sense, as they correspond to perspectives of practically unrealizable "idealized" agents—in our framework, they can best thought of as an "ideal cognitive state". Therefore, connecting our framework more explicitly to these perspectives would require extending it to handling unreachable reward functions, as discussed in Appendix A.9.

G Additional Related Work from Philosophy, Economics, and AI

G.1 Philosophy

Early work on personal identity over time and implications for rational decision-making. While the nature of personal identity under changing selves has been discussed for centuries (Locke, 1689), to the best of our knowledge Parfit (1982) was one of the first to discuss its implications with regards to our conception of rationality. In particular, Parfit describes the challenge with being *timelessly* concerned with evaluating one's life as a whole, without considering time—he rejects this as a practical possibility because of the reality that one inhabits time, and every evaluation comes from the perspective of a particular time. With DR-MDPs, one could say we are trying to explore different approaches for an AI assistant to be timelessly concerned in this way for the wellbeing of a person: while the ultimate goal of finding a "correct" timeless evaluation may be futile, even settling on reasonable evaluations is challenging. The arguments from Parfit (1982) were further expanded in the seminal work "Reasons and Persons" (Parfit, 1984). Of particular note is Parfit's "present aim theory", according to which an individual's rational actions should be guided by their present aims or goals, without giving special weight to their future aims or goals (which seems broadly

[37] Here we assume again that preferences do not change during reward learning time.

equivalent to our initial-reward objective $U_{\text{IR}}(\xi)$),[38] and Parfit's "self-interest theory", which holds that a person should make decisions based on what will be best for them in the long run, even if it conflicts with their present desires or goals (the corresponding objective in our framework seems less clear, as "what is best in the long run" is not specific enough).

Welfare and rationality under changing preferences. Around the same time, Griffin (1986) also criticizes the "totting-up model" of well-bring, according to which we should simply sum well-being over time. There have been many philosophical works focused on the topic of personal welfare in the context of changing preferences: Velleman (1991) considers the relation between the welfare value of a temporal period in someone's life and his welfare at individual moments during that period. Rosati (2013) describes the "narrative thesis", which posits that the way we think of the storyline of our life contributes (in it's own right) to our well being, and echoes related psychological theories (Bauer et al., 2008). Bykvist (2006) states that, for judging inter-temporal decisions, one shouldn't simply look at a single timestep's point of view—we should consider the potential people we could become, and how *they* would evaluate the worlds they are in. Note that this assumes that we trust the assessments and point of view of future selves, which is questionable if we are worried about undue influence. Building on Bykvist (2006), Paul (2014) and Callard (2018) argue that there is no rational basis for making decisions that change the self. Pettigrew (2019) expands on this line of work, and proposes a theory of individual decision making under changing selves based on taking a weighted average between the utility functions of different selves, challenging the notion of "impossibility of grounding rationality" under changing selves. While Paul did not find it convincing (Paul, 2022), we think the framework proposed by Pettigrew (2019) makes significant steps forward. That being said, challenges remain with regards to the details of how weights would be chosen in practice for multi-step decision making: in particular, if weights are re-assessed at every decision making node, it seems even reasonable choices of weights could lead to inconsistent plans. Also, it seems unclear why one could expect a clean separation between one's current weights for different selves, and their current utility function: for instance, wouldn't the utility function of Bob from Figure 1 already include information about how much Bob would want to weigh being a happy conspiracy theorist in the future, as is (implicitly) the case in our example?[39] The main difference between Pettigrew (2019) and our work is that while Pettigrew is squarely focused on human decision making, DR-MDPs can instead be thought of as focusing on choosing objectives that lead to reasonable (and consistent) AI plans, which implicitly account for all relevant selves. Another significant difference is our greater focus on the role of influence, and its implications for the legitimacy of the resulting preferences. On first impression, Pettigrew's framework may seem more flexible, as it accounts for uncertainty in future weights, while DR-MDPs assume that the dynamics of reward functions are known. However, DR-MDPs can still account for uncertainty in future outcomes (and reward functions) using stochastic transitions (even though we don't do this in our examples). For a recent comprehensive review and synthesis of philosophical positions on the problems of changing selves, including a more in-depth summary of Pettigrew's framework, see Strohmaier & Messerli (2024).

Ethics of influence and of nudging. There is a lot of philosophical work on the ethics of influence and manipulation (Noggle, 2020). A specific area closely connected to our issue involves the ethical considerations of "nudging". This concept emerged from behavioral economics and refers to the efforts of institutions to steer the behavior and decision-making of groups towards certain outcomes (Thaler & Sunstein, 2008). Our problem setting is very closely related: in our case, it is an AI system (rather than an institution) that is deciding whether to nudge a user. Perhaps unsurprisingly, the "optimality" and "acceptability" of nudging in the literature are similarly unclear: while nudging was originally promoted as a tool to encourage pro-social outcomes, when it is ethical or overly paternalistic has often been contested (Hansen & Jespersen, 2013; Hausman & Welch, 2009; Thaler, 2018). There are various philosophical works which propose frameworks for external decision-makers to assessing the ethics of whether to perform a specific nudge (Paul & Sunstein, 2019; Pettigrew,

[38] Potentially when used under re-planning, which is discussed in Appendix D.7.

[39] This is related to discussions of how meta-preferences may be easily expressible in our framework if the state were to include enough information to recover θ from it, as discussed in Appendix F.1.

2022). In particular, Paul & Sunstein (2019) claim that a nudge is legitimate if the nudged person is better off, *as judged by themselves* after the nudge. Pettigrew (2022) points out that this heuristic can be misleading, in the case that the nudge was illegitimate (e.g. if it manipulates the person to have different preferences), and proposes a stronger condition as heuristic: that people agree, before and after the nudge, that the nudge was beneficial. Note that the property of Unambiguous Desirability proposed in Appendix E.4 can be thought of as a generalization of the heuristic proposed by Pettigrew (2022), for arbitrary multi-timestep nudges in the form of AI policies.

Work at the intersection of AI and philosophy. Discussions about what forms of influence are appropriate have also taken place at the intersection of philosophy and artificial intelligence: there have been works on value changes (Ammann, 2024), preference change (Kolodny, 2022; Zhi-Xuan et al., 2024), influence (Benn & Lazar, 2022; Bezou-Vrakatseli et al., 2023; Gabriel et al., 2024), and manipulation (Benn & Lazar, 2022; Carroll et al., 2023; Gabriel et al., 2024). The concept of "informed preferences" (Gabriel, 2020), ideal observer theory (Firth, 1952), and Coherent Extrapolated Volition (CEV) (Yudkowsky, 2004) also relate to our work, as we discussed in relation to privileged reward Appendix F.7 (and other appendix discussions, such as that of Appendix B.3). Various works have also tackled the problem of whether it is possible to attribute intent to AI systems actions, and how one could do so (Ward et al., 2024; Halpern & Kleiman-Weiner, 2018). System intent and incentives for influence are closely related (Carroll et al., 2023), and intent has already been studied as a coordination mechanism across time for individual decision-making under changing preferences (Bratman, 1987).

G.2 Economics

Early work on welfare under variable tastes. While others have alluded to changes in tastes and its implications for welfare (Samuelson, 1937; Hayek et al., 1941), to our knowledge Harsanyi (1953) was the first economist that investigated these issues in more depth. In this work, Harsanyi considers a notion of welfare grounded in what people say they prefer rather than in external normative principles such as social welfare. Note that this is a normative stance in its own right, and is similar to how our notions of rewards are grounded in what people say they want. Moreover, unlike our framework, Harsanyi's notion of welfare requires assuming comparability between the utility judgements of a person before and after their preference change. A similar move is taken by von Weizsäcker (1971), who builds on works such as that of Peston (1967). Elster (1979) discusses the limitations of von Weizsäcker's assumptions in more depth, emphasizing the challenges regarding paternalism and which perspectives should be considered to ground welfare judgements when preferences change—which are central to our work. In "Sour Grapes" (Elster, 1983), Elster later also discusses in depth a specific class of preference changes which occur unintentionally to oneself, namely "adaptive preferences". In the same work, Elster also briefly discusses the limitations of state-dependent preference formulations (relevant to Appendix A.5).

Explaining away preference changes as time-inconsistent discounting. Shortly after Harsanyi's work, Strotz (1955) studied the phenomenon of time-inconsistency in human's behaviors. Why do we not save up for retirement enough, and then regret it? Strotz showed that only exponential discounting leads to consistent (re-)planning, therefore people must implicitly not be using that kind of discounting (as they exhibit time-inconsistent behavior). He then discusses two ways that people can account for their time-inconsistency: via commitment devices, or by only considering plans that they would actually be able to follow through on without inconsistency.[40] Note that this stance is implicitly still assuming that people's underlying preferences are static, but just that their discounting scheme is such that they would exhibit time-inconsistent behaviors nonetheless. This move of "explaining away the appearance of preference changes" by casting them as suboptimal discounting with fixed preferences ended up influencing much of the later work on the topic, as reviewed by (Loewenstein et al., 2003). Even though the model of hyperbolic discounting may have sufficient explanatory power of people's decision-making in many settings that economics is interested in (Benzion et al., 1989; Chabris et al., 2008), this is not the case more broadly (Loewenstein & Prelec, 1992; Frederick et al., 2002; Loewenstein et al., 2003).

[40] This analysis was later corrected by Pollak (1968).

Shying away from modeling changing preferences. Why did early economics works—with few exceptions—avoid analyzing *changing* preferences? George (2001) and Grüne-Yanoff & Hansson (2009) give various reasons: preference creation and change have historically been considered topics that lay outside the scope of economics;[41] macroeconomists believed that institutional change (relative to changes in individual's preferences), is by far the more important explanatory factor of economic growth; and maybe most importantly, many microeconomists were of the conviction that human preferences ultimately do not change, and even if they did, it was mathematically counterproductive to model such changes. The most impactful work from this last camp was undoubtedly that of Stigler & Becker (1977). They go as far as to say that "no significant behavior has been illuminated by assumptions of differences in tastes", and that analyses considering changing tastes "give the appearance of considered judgement, yet really have only been ad hoc arguments that disguise analytical failures". Grüne-Yanoff & Hansson (2009) interpret their position as follows:

> "This position may be interpreted either as the ontological claim that preferences indeed are stable, or alternatively as the methodological claim that explanations based on stable preferences are better than those that refer to preference changes. The second interpretation can be based on the assumed relation between explanatory power and simplicity: explaining any conceivable human behaviour through the paradigm of individuals maximizing utility constrained by income and present capital stocks is simpler than supposing that tastes change."

While we agree with the risk of introducing unnecessary formal complexity, we think that in the context of AI interactions with humans, influence effects are too important to be ignored.[42] And as we show in our analysis, ignoring such effects has a cost—that they will likely be optimal under standard objective functions (or notions of welfare, to use the language of economists).

Egonomics, weakness of will, and adaptive preferences. A notable line of work in economics is that of weakness of will and self control problems, introduced by Schelling (1978) with his notion of "egonomics". This area investigates questions of self control and preference change, in particular with respect to internal conflict and addictive behaviors such as smoking. Schelling (1985) discusses the enforcement of rules for oneself, which is relevant to our discussion: people are not always able to commit to the plans that they agree are good for them, often leading them to turn to external accountability partners or enforcers. They play a similar role to the AI systems in our work, as they aggregate across the wishes of the person at different times, and try to encourage their best guess of the behavior that the person would ultimately want for themselves.[43] Our work sidesteps most of these issues around consistency of plans and self-control (Pollak, 1968) by considering an AI assistant's actions, which unlike humans, can credibly commit to carry out a plan (assuming the person cannot switch it off). Importantly, the settings considered by these works are ones in which the normatively desirable course of action is clear, which removes the main complications of dealing with changing preferences: the notion of what rational behavior should consist of remains mostly unthreatened. Concurrently, Elster (1985) interprets weakness of will as a collective action problem between the different selves. For a survey of other works in the literature of weakness of will, see Ainslie (1975).

Recent economics work has started contending with changing preferences more directly. In recent decades, there have been many more works on the topic of changing preferences (Loewenstein et al., 2003; Grüne-Yanoff & Hansson, 2009). In particular, Loewenstein et al. (2003) identifies several sources of preference change: habit formation, satiation, visceral factors, maturation, conditioning, social influences, and motivated taste change. George (2001) formulates an theory of individual welfare that can account for changing preferences by appealing to second-order preferences. To address the regress problem, this work argues that preference changes are most commonly

[41] On this point, see also von Weizsäcker (1971) and Pollak (1978).

[42] For further criticism of Stigler & Becker (1977)'s position, see Pollak (1978).

[43] The usage of "self" in our work comes with similar caveats to those described by Schelling (1984), although we recognize that this usage is vague and arguably misleading—for some criticisms, see how the usage by Schelling was criticized by Elster (1985).

first-order ones, and even when second-order preferences occur, as long as they don't move in tandem with first-order changes, welfare assessments are still possible. Ullmann-Margalit (2006) questions the idea that one could possibly be rational about "big decisions" which change the self, claiming that in a economics sense, there is no footing for a rational choice in these situations, as the "rationality base" changes as a consequence of the decision—setting the groundwork for Paul (2014)'s argument about non-commensurability across different selves. More recently, Bernheim et al. (2019) attempt to model and unify various preference change phenomena under a single descriptive theoretical model, according to which individuals choose their preferences according to what they expect will maximize their utility (subject to their level of "open-mindedness"). The work from Dietrich & List (2013) is also of note: they also propose a descriptive model of preference change, based on "motivationally salient" properties of alternatives available to the agent changing. However, both of these descriptive accounts have yet to be tested (to our knowledge), and it's unclear what the normative implications for decision making of this model should be—which are the focus of our paper.

Unambiguous Desirability and Individual Rationality. The property of unambiguous desirability was inspired by the notion of "individual rationality" from algorithmic game theory (Nisan et al., 2007), which captures the notion of whether any of the individuals involved in an ongoing deal would ever prefer to defect. This is also known as a "participation constraint" or "voluntary participation" (Jackson, 2014).

G.3 AI

Multi-objective MDPs. With a choice of $U(\xi)$, one implicitly replaces the multiple competing notions of optimality (corresponding to each θ) with a single one. The process of choosing a single $U(\xi)$ which implicitly reduces a DR-MDP to an MDP, is similar to the *scalarization* step in Multi-Objective MDPs (Roijers et al., 2013) which reduces a MOMDP to an MDP, which similarly requires an implicit *value judgement* (Chankong & Haimes, 2008). However, DR-MDPs importantly differ from MOMDPs, in that the objectives which should be used to evaluate a trajectory may depend on the trajectory itself (as the actions taken can affect the selves that are realized). Relatedly, MOMDPs don't keep track of which reward function was associated with each step (as it's meaningless in their framing).

Interdisciplinary AI work. The adaptive and changing nature of human feedback has also been emphasized by Lindner & El-Assady (2022). We think there a good area of inspiration for tentative solutions is that of Fiduciary AI (Benthall & Shekman, 2023). More broadly, our conclusion about the challenges of avoiding normative choices when operationalizing alignment rings similar to the points made by Dobbe et al. (2021) and Kirk et al. (2023).

Welfare under conflicting preferences. One of the works which is most related to ours is Kleinberg et al. (2022): in the setting of recommender systems, they also study what should be conceived of as user welfare, focusing on preference conflict. In particular, they model users as having "system 1" and "system 2" preferences that can be in conflict, and show how only optimizing engagement will often be insufficient to guarantee welfare. Our work can be thought of as extending theirs to considering many possible preferences that can change over time, thus modeling change in addition to conflict. Moreover, we don't assume access to any ultimate notion of welfare (which in their case is the judgement of system 2).

Algorithmic amplification in social media, and $\pi_{\mathbf{noop}}$. The study of algorithmic amplification in social media (Thorburn et al., 2022; Ribeiro et al., 2023; Huszár et al., 2021) can be thought of as a study of influence emerging from specific algorithmic choices. Notions of amplification also need to be specified relative to a "neutral" baseline, similarly to our notions of influence (Definition 7): what baseline is most appropriate has been subject to debate (Laufer & Nissenbaum, 2023; Thorburn, 2023). Maybe the most common choice in practice has been comparing to a reverse chronological recommender (Huszár et al., 2021; Milli et al., 2023), but others include comparing to others recommender systems (Fast et al., 2023), no platform at all (Allcott et al., 2020), or random recommendations (Carroll et al., 2022).

Performative prediction, performative power, and preference influence. An interesting line of work which has emerged in recent years is that of performative prediction (Perdomo et al., 2020) and performative power (Hardt et al., 2022), which concerns itself with the capacity of classifiers to affect the distribution of their future inputs. This idea is similar in spirit to the work of Krueger et al. (2020), and is connected to our concern with AI systems' capacity to influence humans. However, we see various reasons to prefer the RL formalism to that of Hardt et al. (2022) for the types of influences were are interested in: performative prediction and power are mostly focused on firms which operate in sequential decision problems (e.g. domains in which the algorithm's choices affect future users' behavior), but use algorithms that myopically optimize over only the next timestep's outcomes—from this perspective, they only allow to model 1 of the 8 objectives we consider in Table 2. For instance, to the best of our understanding, performative power (Hardt et al., 2022) can be thought of as a measure of how much a firm can shift users over the course of *a single timestep*, if they choose to do so. The steering analysis of ex-ante and ex-post optimization only performs a one-timestep lookahead, feels like a less natural formalism for the multi-timestep nature of most preference changes—especially if one considers that the RL formalism solves the multi-timestep generalization of the ex-post optimization problem by design: in RL training, the human's adaptation to the AI is already factored into how the AI should be making decisions in order to maximize the multi-timestep objectives. In short, the lens of RL seems strictly more expressive and more suited to our purposes than that of performative prediction, but comes at the cost of additional computational challenges. As a final point of comparison, the framing of Hardt et al. (2022) is mostly focused on the misalignment between firms and targets of the firm's algorithms—focusing on the power that firms have to steer them to their benefit (under well-defined metrics). Instead, we can be thought of as focusing on the problem of "steering oneself" with the help of an AI system, and the challenges which remain in determining an appropriate metric for success, as discussed in relation to "egonomics" (Schelling, 1978). Essentially, while we recognize that firm-user misalignment is a very significant reason for worry, we focus on the challenges that would remain even if AI systems were to be developed solely with user alignment in mind.

Social choice theory. Our setting shares various similarities with preference aggregations across multiple individuals, as mentioned in Section 1. The problem of social aggregation is studied by social choice theory (List, 2022; Brandt et al., 2012), which mainly differs from our framework in that there is no temporal dependency between elements of the decision which is being made. There have also been recent works discussing the parallels between social choice and AI Alignment (Mishra, 2023; Conitzer et al., 2024), in particular with respect to RLHF. While there has been some work focusing on collective decision-making across time (for individuals who can change their preferences), these works mostly ignore the influence incentives which emerge from their notions of optimality (Parkes & Procaccia, 2013; Freeman et al., 2017; Kulkarni & Neth, 2020). Indeed, to our understanding, the objectives proposed by these works could lead to undesirable influence incentives: for example, in Parkes & Procaccia (2013) the optimal social choice function (i.e. policy for the social choice MDP) may initially take actions dispreferred by everyone if this can influence future preferences to be in accordance with one another (guaranteeing future unanimity). An interesting direction for future work would be to consider traditional social welfare functions (e.g. Rawlsian, Nash, Utilitarian, etc.) as objectives for the aggregations across the different selves. However, in some preliminary investigations in this direction, it seemed to us that such approach would not necessarily achieve more desirable influence properties than the objectives we considered in Table 2.

Reward uncertainty in MDPs. There have been some works accounting for uncertainty in reward functions: most relevant is the work by Regan & Boutilier (2010), who consider Imprecise Reward MDPs, in which there is a feasible set of reward functions that cannot be disambiguated by, e.g., acting in the environment. For such setting, they define the optimization objective in terms of minimax regret, that is, they aim to minimize the maximum regret incurred if the worst reward function from the feasible set were to be chosen. We could have also included this criterion in our analysis and in Table 2, but we are confident it would also run into issues—minimax regret in this setting would require comparing regret across different possible reward functions, requiring "interpersonal" comparisons that could always be set to be in favor of undesirable manipulation incentives. Other works also account for reward uncertainty, but assume that there is one true Markovian reward function that one can learn about, and one should use to evaluate trajectories (Hadfield-Menell et al., 2016; Desai, 2017; Chan et al., 2019). A gap in the literature is work which tries to disambiguate between reward changes, and updates in belief regarding reward functions (which we discuss briefly in Appendix A.3).

H Limitations and Discussion

Learning Θ and its dynamics. Throughout the paper, we assumed that the human reward functions and their dynamics were known. In practice, they would have to be learned—which would reward learning techniques that account for reward dynamics, and committing to a choice of what counts as a "cognitive state" θ relative to the external state s (Appendix A.6). However, such limitation strengthens our conclusion: it shows that even with full knowledge of human (stated) preferences, there seems to be no neat resolution for the normative challenges which arise—indicating that this difficulty is inherent to handling preference changes themselves, rather than an artefact of uncertainty over them.

Existence of $a_{\textbf{noop}}$. Similarly to other AI safety work (Krakovna et al., 2019; Farquhar et al., 2022), some of our definitions assume there is a a_{noop} action (and π_{noop}). Despite the challenges in grounding notions of π_{noop} in practice, we think that formulating our notion of influence in terms of π_{noop} can still be helpful for theoretically analyzing the properties of AI systems. If it is especially hard to operationalize π_{noop} for a given setting, it seems all the more important to be cautious about possibilities for influence in such a domain. Approximating notions of π_{noop} has been attempted by various prior works (Carroll et al., 2022; Huszár et al., 2021), and has been discussed extensively in the "algorithmic amplification" literature (Appendix G.3).

Unreachable θs, and meta-preferences. To simplify our analysis, we restrict our analysis to reachable cognitive states. However, cognitive states which we aspire to may *not* be reachable in practice (Firth, 1952; Yudkowsky, 2004). Accounting for non-reachable reward functions would require additional complexity, which would only increase the need for challenging normative judgements (see Appendix A.9). Another limitation of our formalism is that it can only express preferences for AI influence implicitly (by having influence be optimal under a human's reward function). Having such meta-preferences (George, 2001; Franklin et al., 2022) be expressible explicitly, i.e. allowing reward functions to *evaluate transitions between different reward functions*, may be useful to more clearly capture notions of legitimacy of influence and personal autonomy.

Simplicity of our examples. While our example DR-MDPs are simple, they are sufficient for our purposes: they provide proofs of existence of failure cases for each of the objectives we consider (by being overly conservative or leading to undesirable influence). That being said, extending analyses of influence incentives to more realistic settings with real human data is an important direction for future work in developing agents that behave acceptably in practice.

Misaligned economic pressures. We show that even if AI systems were solely designed with users' welfare in mind, is it not clear how to avoid undesirable influence without while retaining capabilities. As real-world AI systems will instead be developed under strong economic incentives that will often be at odds with users' well-being (Susser et al., 2018), this gives additional reason for worry.

Changing societal values and norms. The DR-MDP model seems easily adaptable to multi-agent settings in which the changing reward functions correspond to the changing aggregated preferences of a collective, rather than the preferences of a single human. This may be a fruitful avenue of investigation, which expands on prior work from social choice (Parkes & Procaccia, 2013).

I Broader Impacts

That people are influenceable is already known and leveraged in practice by relevant industries. The fact that people are influenceable is already well established and leveraged by many industries of mixed ethical connotations: marketing and advertisements (Pickett-Baker & Ozaki, 2008; Ayanwale et al., 2005), political campaigning (Wylie, 2020), therapy (Moyers & Martin, 2006), education (Hyman & Wright, 1979), and government nudging units (Halpern & Sanders, 2016). Optimizing for specific influence outcomes is simple even with the current AI paradigm, and is already being done in practice, such as with the case of engagement (Irvine et al., 2023; Cai et al., 2023), purchases (Gauci et al., 2019), improving educational outcomes (Bassen et al., 2020), or improving emotional well-being (Cunningham et al., 2024).

Not modeling influenceablity does more harm than good. *Not* modelling the problem of preference change is *not* a solution, in ways that shares parallels with the limitations of fairness through unawareness (Dwork et al., 2011; Teodorescu, 2019) or security through obscurity (Moshirnia, 2017). We argue that most real-world systems trained and deployed with humans will affect our preferences, regardless of whether it is our intention or not. To mitigate issues that may arise from this, we require a model that explicitly accounts for preference change. This work takes a step in that direction, without providing any additional means by which malicious actors may act (to our current understanding).

Our work does not meaningfully increase the capacity to influence people. In this work, we have detailed a framework that models preference change within an MDP-like model. While one may be able to leverage our theoretical insights to make systems more capable of influence, attempting to influence in targeted ways is already quite straightforward without our formalism (e.g. just rewarding the system for desired influence outcomes). Instead, it's much more challenging to have the system *avoid* inducing undesired forms of influence as side-effects, which is the main focus of our paper.

With these points in mind, we believe that our work does not pose a societal risk, but rather allow to explore questions which will be fundamental to address for the ethical deployment of AI systems.