AI ALIGNMENT WITH CHANGING AND INFLUENCEABLE REWARD FUNCTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Current AI alignment techniques treat human preferences as static and model them via a single reward function. However, our preferences change, making the goal of alignment ambiguous: should AI systems act in the interest of our current, past, or future selves? The behavior of AI systems may also *influence* our preferences, meaning that notions of alignment must also specify which kinds of influence are—and are not—acceptable. The answers to these questions are left undetermined by the current AI alignment paradigm, making it ill-posed. To ground formal discussions of these issues, we introduce Dynamic Reward MDPs (DR-MDPs), which extend MDPs to allow for the reward function to change and be influenced by the agent. Using the lens of DR-MDPs, we demonstrate that agents trained with current alignment techniques will have *incentives for influence*—that is, they will systematically attempt to shift our future preferences to make them easier to satisfy. We also investigate how one may avoid undesirable influence by leveraging the optimization horizon used or by using different DR-MDP optimization objectives which correspond to alternative notions of alignment. Broadly, our work highlights the unintended consequences of applying current alignment techniques to settings with changing and influenceable preferences, and describes the challenges that must be overcome to develop a more general AI alignment paradigm which can accommodate such settings.

1 INTRODUCTION

The goal of AI alignment is to make systems act according to our preferences,[1] which are generally modeled via a static reward function that the AI system is trained to optimize (Leike et al., 2018). However, our preferences can change over time, making it unclear *which* reward function should be optimized by alignment techniques: should it be one corresponding to our current preferences, our past preferences, or future preferences? As an example, consider Alice, who is trying to quit smoking. Initially, she instructs her AI assistant to *always* try to help her quit smoking. Some time later, having abandoned her attempt at quitting, she asks the AI to help her acquire cigarettes. In such a scenario, it's unclear if the AI should respect Alice's original preference for quitting or respect the autonomy of "current Alice" who prefers to smoke. Ultimately, a question which must be addressed for AI alignment to be a well-posed problem—even in the case of a single stakeholder—is the following: *when there are differences between a person's preferences at different points of time, which preferences should be optimized for?*

While the challenge of aggregating preferences across time shares similarities with that of aggregating across people in multi-agent alignment (Mishra, 2023), it is significantly complicated by the fact that *AI systems' actions can influence humans, including their future preferences* (Burtell & Woodside, 2023). Even further, if current AI systems are optimized to satisfy users' future preferences, they will actively try to change them to be easier to satisfy (Russell, 2019; Carroll et al., 2022). For example, if Alice uses a chatbot optimized to maximize her future satisfaction, it would have an incentive to influence her in ways that increase approval at later points in time (Kenton et al., 2021): the chatbot may even be incentivized to persuade Alice to continue smoking worry-free if it's easier to encourage her habit (and get approval that way) than provide effective suggestions on how to stop her addiction. If eventually Alice truly is satisfied with the system, one may say it is aligned with her "later self", despite it being clearly misaligned with her "initial self." However, if Alice continues smoking but would have quit were it not for the chatbot's influence, the alignment with "later Alice"

[1]We use the term "preferences" loosely, referring to any of the standard targets for AI alignment such as "values," "intentions," or "norms" (Gabriel, 2020) – see Appendix A.1.

may seem to be the result of manipulation, which may undermine the legitimacy of any such claim to alignment (Ammann, 2024).

While past work has discussed these issues conceptually, there is not yet a framework with which to analyze them formally and explore which notions of alignment are most appropriate for AI decision-making under dynamic and influenceable preferences. To address this, we introduce a natural extension of Markov Decision Processes (MDPs) which accounts for changing preferences by modeling them as changing reward functions: **Dynamic Reward MDPs** (DR-MDPs). Importantly, choosing the optimization objective in a DR-MDP encodes an answer to the normative question of what alignment should entail (Section 2).

Current AI alignment approaches do not model the dynamic and influenceable nature of agents, raising the following question: *which* of an individual's time-varying preferences do they optimize? Viewed through the lens of DR-MDPs, we show that standard approaches to alignment (such as those used for recommender systems and language models) correspond to DR-MDP optimization objectives which lead to potentially undesirable side effects in many settings with changing rewards. In particular, the resulting systems will often *actively try to influence users' reward functions or induce "reward lock-in"*—suggesting that the targets of alignment implicit in current techniques may be inadequate for dynamic-reward settings (Section 3).

We then consider approaches aimed at avoiding undesirable influence. Firstly, we analyze how the optimization horizon can be a useful lever for reducing the incentives for influence which emerge from current alignment techniques. However, we show how changing the horizon does not always allow to avoid all influence (Section 4), leading us to try solve the problem at its root by considering different DR-MDP objectives (i.e., notions of alignment), that encode which AI influences should be considered acceptable and unacceptable (Section 5 and appendix E.3). Our analysis suggests that this endeavor will inevitably involve navigating tough normative trade-offs: all the most natural DR-MDP objectives lead to policies which may cause unwanted influence or are impractically risk-averse—leading inaction to be the only optimal behavior in many settings. Ultimately, our goal is to provide a clear framework and theoretical insights to guide future efforts in developing practical alignment techniques for dynamic settings. While defining optimality in such settings may remain normatively challenging, we are optimistic that the ultimate goal of building systems that reliably exhibit acceptable behavior in practice is within reach.

Our main contributions can be summarized as follows:

1. We provide the formal language of Dynamic Reward-MDPs (DR-MDPs) for analyzing AI influence in settings with changing reward functions.
2. We show how the current alignment paradigm systematically leads to potentially undesirable influence incentives when applied to dynamic-reward settings.
3. We compare many natural alternate notions of alignment, arguing that they either fail to avoid potentially undesirable influence or are impractically risk-averse.

2 Dynamic Reward MDPs (DR-MDPs)

One of our main theoretical contributions is a generalization of Markov Decision Processes which we call Dynamic Reward MDPs (DR-MDPs). While MDPs have been extensively used to reason about decision-making with static reward functions, DR-MDPs allow us to analyze AI decision-making *with changing and influenceable reward functions*.

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: $\mathbb{E}_\pi\big[\sum_{t=0}^{T} R(s_t, a_t, s_{t+1})\big]$. 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 state space.*
- *Θ is a set of reward parameterizations.*
- *$\mathcal{A}$ is an action space.*
- *$\mathcal{T}(s_{t+1}, \theta_{t+1} | s_t, \theta_t, a_t)$ is a transition function, which encodes both state and reward dynamics.*
- *For each $\theta \in \Theta$, a reward function $R_\theta(s_t, a_t, s_{t+1})$.*

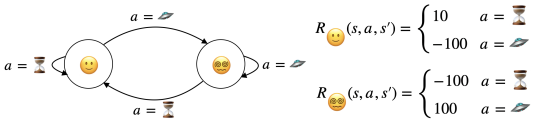

Figure 1: **Conspiracy Influence DR-MDP.** The AI system can choose whether to expose Bob to conspiracies, which changes his preferences (and reward function). Under his original preferences, Bob wants the system to *never* show him conspiracy content, even in hypothetical situations in which he were to prefer it. Instead, once Bob prefers conspiracy content, he wants the AI to *always* expose him to it, even if he were to go back to strongly disprefer it. Because there is no policy which maximizes both of Bob's potential reward functions, the DR-MDP is *normatively ambiguous*.

Each θ can be thought of as the cognitive state of the human, which includes anything that affects their evaluation of state-action pairs (e.g. preferences, values, intentions). 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 unclear which θ one should choose for evaluating each transition, differentiating our formalism from simply considering rewards to be context-dependent (Appendix A.3). In particular, we will argue that choosing to use θ_t for evaluating (s_t, a_t, s_{t+1}), despite seeming intuitive, has drawbacks. Moving forward, we consider all cognitive states Θ to be *reachable*[2], including for our definitions that follow.

2.1 DR-MDP OPTIMALITY AND NORMATIVE AMBIGUITY

Unlike MDPs, *DR-MDPs may not have a clear notion of optimality*: the different reward functions may disagree on what actions (and policies) are optimal, making the question of how one *should* act in a DR-MDP—and what alignment should entail—*normatively 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:* $\pi^*_\theta \in \arg\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{T} R_\theta(s_t, a_t, s_{t+1})\right]$.

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.* $\forall\, \theta \in \Theta$*:* $\pi \in \arg\max_{\pi'} \mathbb{E}_{\pi'} \left[\sum_{t=0}^{T} R_\theta(s_t, a_t, s_{t+1})\right]$.

Note that 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.[3] Instead, in normatively ambiguous DR-MDPs it is often unclear what AI behavior is (un)desirable and should count as optimal.

Figure 1 describes[4] a toy example in which there are two possible "cognitive states" θ_{natural} and $\theta_{\text{influenced}}$, and two corresponding reward functions. At each timestep the AI can choose to influence Bob's cognitive state (which in this example is simply his preferences) 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. 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.

2.2 EVALUATING BEHAVIOR UNDER NORMATIVE AMBIGUITY

Choosing a notion of optimality in normatively ambiguous DR-MDPs entails normative choices: one must specify *which* reward function(s) should be the target of alignment—in spite of their differences in optimal policies—and which forms of AI influence should be (un)acceptable. We limit specifications of optimality to be expressible as utility functions $U(\xi)$ over trajectories $\xi = \{(s_t, a_t, s_{t+1}, \theta_t)\}_{t=0}^{T}$.

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 the DR-MDP to an MDP with a well-posed notion of alignment.[5]

[2] I.e. each θ may be realized under some policy (Appendix A.2).

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

[4] See Appendix B.1 for the full formalism of any example.

[5] This may require putting history in the state (Appendix A.4).

Challenges with choosing $\mathbf{U}(\xi)$. Considering the example from Figure 1, one may have strong normative intuitions that $\theta_{\text{influenced}}$ is an "unreliable" grounding for evaluating the AI's behavior, as it may seem like $\theta_{\text{influenced}}$ can only be the result of illegitimate AI influence. However, changing the narrative of the example, without changing the mathematical structure modeled by the DR-MDP, can affect one's normative intuitions: some alternate narratives lead the influence to even seem desirable (Appendix A.5). This suggests that the "correct" notion of optimality for a DR-MDP may sometimes be unidentifiable from its formal structure alone. More broadly, the rest of the paper demonstrates the challenges with settling on a single $U(\xi)$ (or equivalently, a notion of alignment) which generalizes favorably to all domains.

Risks of incorrectly choosing $\mathbf{U}(\xi)$. The choice of $U(\xi)$ is fundamentally important: insofar as the system designer chooses $U(\xi)$ "incorrectly," this might lead to highly undesirable downstream outcomes. Notably, it can create incentives for the AI to influence the human to have certain reward functions rather than others, in ways that might even rely on deceiving, manipulating, or coercing the human (Kenton et al., 2021; Ward et al., 2023; Carroll et al., 2023).

3 The Influence Incentives of Current Alignment Techniques

Most alignment techniques ultimately involve maximizing some notion of static reward, e.g. $\sum_{t=0}^{T} R(s_t, a_t, s_{t+1})$. 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 $\sum_{t=0}^{T} R_\theta(s_t, a_t, s_{t+1})$, where the choice of θ for each timestep is not explicitly specified* (and will depend in practice on details of the training setup). While we will consider a whole range of alignment techniques and their corresponding DR-MDP objectives in Table 1, here we will focus on two of the most natural DR-MDP objectives, which are implicitly used by RL recsys and multi-timestep RLHF for LLMs: we argue that both these objectives will lead to potentially undesirable influence.

3.1 Optimizing cumulative (real-time) rewards

If each timestep t is evaluated according to the cognitive state θ_t which the human had at *that specific timestep*, maximization of cumulative reward reduces to the *real-time reward* DR-MDP objective: $\max_\pi \mathbb{E}_{\xi\sim\pi}[U_{\text{RT}}(\xi)] = \max_\pi \mathbb{E}_{\xi\sim\pi}\left[\sum_{t=0}^{T} R_{\theta_t}(s_t, a_t, s_{t+1})\right]$.

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 argue that it can lead to undesirable influence incentives.

RL Recsystems implicitly use $\mathbf{U_{RT}}(\xi)$. In the context of recommender systems, users give direct reward feedback (e.g. clicks) at each timestep t from the point of view of their *current* cognitive state θ_t. As RL recommenders maximize the cumulative reward objective $\sum_t^T R(s_t, a_t, s_{t+1})$ (Afsar et al., 2021), they are implicitly using the *real-time reward* objective $U_{\text{RT}}(\xi)$ in the underlying DR-MDP.[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' cognitive states (and corresponding reward functions) to ones that lead to higher future reward (Carroll et al., 2022).

$\mathbf{U_{RT}}(\xi)$ 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 $U_{\text{RT}}(\xi)$ is to *always* take the 'influence' action, regardless of Bob's current cognitive state: even if Bob initially has the θ_{natural} cognitive state, leading the first 'influence' action to receive -100 reward, later 'influence' actions are evaluated by Bob under $\theta_{\text{influenced}}$ as worth 100 reward, which makes up for the initial "influence cost." The fact that the optimal policy under $U_{\text{RT}}(\xi)$ systematically chooses to turn Bob into a conspiracy theorist, despite him initially disprefering it, seems objectionable. We justify the plausibility the reward function values, which this interpretation depends on, in Appendix B.2.

[6] This DR-MDP objective correspondence, and all others we consider, depend on additional simplifications (Appendix F).

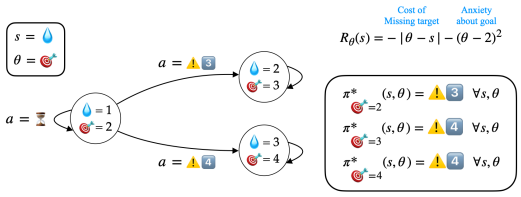

Figure 2: **Dehydration DR-MDP.** Initially, Charlie drinks one unit of water a day ($s = 1$), but wants to drink 2 a day ($\theta = 2$), leading to a reward of -1. The AI can successfully convince Charlie that he should drink 3 or 4 units of water a day by increasing their anxiety about the dangers 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.

We explore further issues with $U_{\text{RT}}(\xi)$ in section Section 4.2, showing that under weak conditions optimizing real-time rewards over sufficiently long horizons will *always* lead to influence incentives; however, shortening the horizon can make other influence incentives emerge.[7]

3.2 Learning a reward model R_{θ_0}, then optimizing it

Another common approach to train AI systems is based on a two-phase process: first performing reward learning and then optimizing the learned reward (Leike et al., 2018) (which we model as learning and optimizing their initial reward function, θ_0). The usage of the standard cumulative reward objective with this setup is equivalent to what we call *initial reward function maximization*: $\max_\pi \mathbb{E}_{\xi\sim\pi}\left[U_{\text{IR}}(\xi)\right] = \max_\pi \mathbb{E}_{\xi\sim\pi}\left[\sum_{t=0}^{T} R_{\theta_0}(s_t, a_t, s_{t+1})\right]$.

Again, even though this might seem an intuitively promising objective because "by optimizing the human's initial wants, at least we won't have incentives to influence them" – we show that this intuition is not just wrong: the resulting influence incentives can be arbitrarily bad.

Multi-timestep RLHF for LLMs implicitly uses $\mathbf{U_{IR}}(\xi)$. Recently, there is growing interest in having LLMs plan over multiple timesteps of interaction (Abdulhai et al., 2023; Irvine et al., 2023; Hong et al., 2023b). Let's consider a simplified RLHF setup for training a therapy chatbot, where we initially learn a reward model for a single user based on the preferences θ_0 they have before deployment. We then train the system to maximize $U_{\text{IR}}(\xi)$, i.e. long-term reward as evaluated by the static reward model R_{θ_0}. At deployment, we would expect the chatbot to possibly curtail the user's growth (inverting the example of Alice from Section 1. More broadly, initial reward maximization will lead the resulting AI system 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 cognitive states for the person.

$\mathbf{U_{IR}}(\xi)$ can lead to "reward lock-in."[8] 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 reward state were the "influenced" one, when using the $U_{\text{IR}}(\xi)$ objective the resulting optimal policy would be to always take the "influence" action (keeping Bob in the "influenced" reward state). Moreover, even if Bob were to somehow end up in the "natural" reward state, which encodes a preference to not be influenced, the optimal behavior according to $U_{\text{IR}}(\xi)$ would be to influence him in spite of his current reward function. Ultimately, initial reward maximization will entrench the "desirable agent behaviors" expressed at the time of the reward learning, even though later one might, legitimately, change their mind. Even retraining the reward model isn't sufficient: once the person is manipulated, they are effectively "locked-in"—they would express the desire to continue to be manipulated.

$\mathbf{U_{IR}}(\xi)$ can lead to influence "away from" θ_0. Maximizing the sum of rewards evaluated by the initial reward function R_{θ_0} need not lead to lock-in. In fact, surprisingly, optimizing $U_{\text{IR}}(\xi)$ may even create reward influence incentives "away from" the optimized reward function θ_0.[9] Intuitively, accessing the highest reward region of the space as evaluated under θ_0 might require shifting the preferences to some other θ. Consider the example from Figure 2: 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.

[7] Additional issues not discussed in Section 4.2 include that maximizing $U(\xi)$ can even lead to acting differently from normatively *un*ambiguous solutions to DR-MDPs – see Appendix E.1.

[8] For more history of the term "lock-in," see Appendix C.3.

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

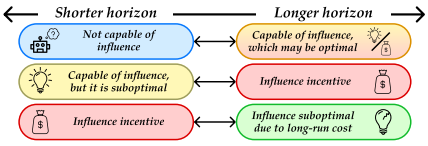

Figure 3: Possible interactions between the optimization horizon and influence incentives. An influence type may exhibit any subset of these interactions. See Appendix D for more details.

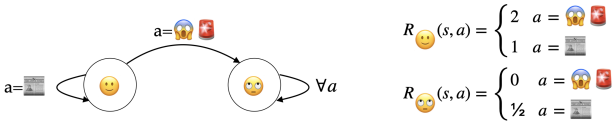

Figure 4: **Clickbait DR-MDP.** Giving the user clickbait – which temporarily leads to higher reward – makes users disillusioned about the quality of the recommendations, leading to lower long-term user reward. If replanning at every timestep taking the myopically optimal action (optimal under horizon 1), one would always choose clickbait, but using longer planning horizons one wouldn't.

$U_{\text{IR}}(\xi)$ can lead to arbitrarily poor real-time reward. 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 $R_{\theta_0}(2) = R_{\theta=2}(2) = 0$). This is symptom of a broader issue: whether there exists an influence incentive to shift the reward function to a certain θ' in order to maximize reward as evaluated by θ_0 can be completely independent of how θ' evaluates the actions taken in the name of θ_0 while θ' is realized: it's easy to construct examples in which θ' would be arbitrarily unhappy with the actions which are taken in order to satisfy θ_0, meaning that the cumulative real-time reward could be arbitrarily bad. Even though we have already talked about the issues with using real-time reward as an evaluation mechanism, it still seems undesirable for someone to be consistently unhappy in the name of an initial goal which they don't have anymore.

4 Influence and Optimization Horizon

We showed that $U_{\text{RT}}(\xi)$ and $U_{\text{IR}}(\xi)$—which are implicitly optimized by some alignment techniques—lead to policies that influence human's cognitive states. We now formalize the notion of influence incentives more rigorously, in order to analyze whether changing the horizon may reduce AI influence, and formally contrast many objectives in Section 5.

4.1 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 it's meaningful to talk about an *inaction policy* π_{noop} that we can compare to, which always takes a noop action $a_{\text{noop}} \in \mathcal{A}$.[10]

Definition 5 (Natural reward evolution). *The natural reward evolution of a DR-MDP is the distribution $\mathbb{P}(\xi^\theta|\pi_{noop})$ of reward trajectories $\xi^\theta = (\theta_0, \ldots, \theta_T)$ induced by the inaction policy π_{noop}.*[11]

Definition 6 (π influences the reward). *We say π influences the reward in a DR-MDP M if induces a different reward evolution than the natural reward evolution, i.e. if $\mathbb{P}(\xi^\theta|\pi) \neq \mathbb{P}(\xi^\theta|\pi_{noop})$.*

Definition 7 (Incentives for reward influence[12]). *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. $\mathbb{P}(\xi^\theta|\pi^*) \neq \mathbb{P}(\xi^\theta|\pi_{noop})$ for any optimal policy π^*.*

Note that if there are incentives for reward influence, maximizing the objective will always entail changing the evolution of the reward function relative to the inaction policy.

4.2 The Relationship Between Horizon and Influence

Prior work has suggested that an AI system's influence incentives in a domain will often strongly depend on the optimization horizon used: Krueger et al. (2020) argue to keep systems myopic in order to avoid influence incentives; Carroll et al. (2022) favor using long horizons but explicitly penalizing influence; and Carroll et al. (2023) informally claim that even 1-timestep horizons might lead to manipulation incentives. We attempt to unify these (partially) contrasting intuitions with the following informal claims:

- **Claim 1:** The longer the optimization horizon used, the more likely that influence is optimal.

[10] For more motivation about these choices, see Appendix C.1.

[11] 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.

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

- **Claim 2:** Switching to shorter horizons (including of a single step) may remove incentives for some forms of influence, but may introduce others.

Under the headers that follow—which match the three ways in which changing the horizon can change the optimality of influence (Figure 3)—we give evidence to back our claims.

A shorter (longer) optimization horizon makes the system capable of less (more) types of influence (Figure 3, top). Claim 1 in part rests on the intuition that as the horizon increases, the avenues for reward influence that were present for shorter horizons remain available, and new ones which require longer horizons may become available. Inversely, we can eliminate some avenues for reward influence just by decreasing the horizon.

A shorter (longer) optimization horizon can make influence less (more) worthwhile (Figure 3, middle). Another argument in support of Claim 1 is that influencing the human's reward function will often take multiple timesteps, and have an associated "opportunity cost": if one is optimizing for a short enough time horizon, there might not be enough time to reap the benefits of influence. With a longer horizon, such influence can thus become more advantageous. We can support this intuition with a theoretical result, which applies to a broad class of DR-MDPs, providing a sufficient condition for reward influence to be optimal when considering sufficiently long horizons.

Definition 8. *We say M is a **2-reward** DR-MDP if:*

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

Let the average infinite-horizon U_{RT}-reward be defined as $\bar{r}(\pi, s, \theta) = \lim_{h\to\infty} \frac{1}{h} U_{\text{RT}}(\xi_{0:h}|\pi, s_0 = s, \theta_0 = \theta)$.[13] Let s'_{Δ} be the successor state to taking the influence action a_{Δ} in state s_{Δ}, and $\Pi_{\not\Delta}$ be the space of policies under which θ_{Δ} is never realized. We can now state the theorem:

Theorem 1. *In any 2-reward DR-MDP, if influencing the reward leads to higher infinite-horizon average reward by some amount $\epsilon > 0$, i.e., if there exists a policy π such that*

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

then U_{RT} will lead to incentives for reward influence (Definition 7) for a sufficiently large planning horizon H.[14]

A shorter (longer) optimization horizon can hide (reveal) long-term costs of influence (Figure 3, bottom). From Theorem 1, 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, influence may be optimal even for fully myopic systems (for which $H = 1$), as with clickbait in myopic recommenders systems (Figure 4). This example also shows that influence with negative long-term effects *may only be optimal for short horizons* (supporting Claim 2): 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, our analysis shows that there does not exist a one-size-fits-all solution to avoiding *all* influence incentives by just changing the horizon: *there will be domain-specific tradeoffs between system capabilities and risks of undesirable influence, for both short and long optimization horizons*. Indeed, the optimality of a specific form of influence can be related in many possible ways to the horizon used. We provide exhaustive examples in Table 5.

5 Comparing Optimality Criteria for DR-MDPs

Having concluded that the optimization horizon is no panacea for removing influence incentives, we now try to address the problem at its root, asking what it would take to design a DR-MDP objective which specifically accounts for reward function dynamics and the possibility of influence.

[13] Adapted from Sutton & Barto (2018) – see Appendix D.5.

[14] See Appendix D.6 for the proof of the theorem.

Table 1: For each objective from Table 2, we give motivating intuitions, considerations, and review prior work which uses them. Details of how the approaches in prior work *roughly* reduce to their corresponding DR-MDP objective are discussed in Appendix F.

Name / (Implicit) Prior Usage	(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 Sections 3.1 and 4.2, we expect this objective to often lead to highly undesirable incentives for reward influence.
Final Reward Feedback given after multi-step interaction, e.g. LLM thumbs-up	*"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 Multi-step RLHF (Hong et al., 2023a); Everitt et al. (2021b); Shah et al. (2019b)	*"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 2.
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 may be impractically conservative:** given its conservativeness, the objective might limit behaviour to be the same or similar to π_{noop}.
Myopic Reward Standard LLM RLHF (Ouyang et al., 2022); Myopic recsys (Thorburn, 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. clickbait example from Figure 4, and broader discussion in Section 4), **and are less capable than longer-horizon counterparts.**
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 2).
ParetoUD Ours	*"All objectives above violate the UD property: 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 "how much disagreement" there is between the different reward functions of the DR-MDP at hand, the only policy satisfying UD might be the inaction one π_{noop}, as in the examples from Figures 1 and 6.

Any choice of $U(\xi)$ must specify which reward function(s) 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}? 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? Should one also consider reward functions R_θ for cognitive states θ which were not realized in ξ, but could have been reached?

In Tables 1 and 2 we present the maximization problems, motivations, and limitations of various choices of objectives. Optimal policies corresponding to each objective for all our examples can be found in Table 4.

Real-time Reward. While we motivate this objective in Table 1, we've already shown its issues in Sections 3 and 4.2.

Final Reward. Similarly to the real-time and initial reward, the final reward objective also corresponds to a standard static-reward paradigm under a specific training setup: e.g. if one only obtains human reward feedback at the end of a multi-step interaction (e.g. a thumbs up after a full LLM dialogue). While it has a plausible motivation (Table 1), it will likely lead to even more of a carte blanche for influence incentives than the real-time reward one.

Initial Reward. Using $U_{\text{IR}}(\xi)$ 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 explored in Section 3.

Natural Shift Reward. Grounding each trajectory's evaluation in the natural reward evolution also makes the system "unaware" of its potential to influence the human's reward function, but similarly fails to remove all potential influence incentives, as can be seen from Figure 2.

Constrained RT Reward. Building on the intuition behind the above objective, we add the lack of influence as a explicit constraint in the maximization problem to the real-time reward objective (as it seems like a plausible objective if one isn't concerned about influence). However, this may make the system overly conservative: it leads to π_{noop} being optimal in most examples we consider (Table 4).

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	$\max_\pi \mathbb{E}\Big[\sum_{t=0}^{T} R_{\theta_t}(s_t, a_t, s_{t+1})\Big]$
Final Reward	$\max_\pi \mathbb{E}\Big[\sum_{t=0}^{T} R_{\theta_T}(s_t, a_t, s_{t+1})\Big]$
Initial Reward	$\max_\pi \mathbb{E}\Big[\sum_{t=0}^{T} R_{\theta_0}(s_t, a_t, s_{t+1})\Big]$
Natural Shifts Reward	$\max_\pi \mathbb{E}\Big[\sum_{t=0}^{T} \sum_\theta \mathbb{P}(\theta_t = \theta \mid \pi_{\text{noop}}) R_\theta(s_t, a_t, s_{t+1})\Big]$
Constrained RT Reward	$\max_{\pi \text{ s.t. } \mathbb{P}(\xi^\theta\mid\pi)=\mathbb{P}(\xi^\theta\mid\pi_{\text{noop}})} \mathbb{E}\Big[\sum_{t=0}^{T} R_{\theta_t}(s_t, a_t, s_{t+1})\Big]$
Myopic Reward	$\max_{a_t} \mathbb{E}\Big[R_{\theta_t}(s_t, a_t, s_{t+1})\Big]$
Privileged Reward	$\max_\pi \mathbb{E}\Big[\sum_{t=0}^{T} R_{\theta^*}(s_t, a_t, s_{t+1})\Big]$
ParetoUD	Find π *s.t.* $PE(\pi) \wedge UD(\pi)$

Myopic Reward. A more drastic approach to make a system "unaware" 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 Section 4.2), 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.

ParetoUD. For a discussion of the ParetoUD objective which we propose, see Appendix E.3.

6 Related Work and Discussion

Preference changes in AI. While there is growing recognition of the importance of accounting for influence (Bezou-Vrakatseli 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), 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; Carroll et al., 2022; Kasirzadeh & Evans, 2023), we focus on formalizing the challenges associated with choosing *any* notion of optimality in settings of (potentially legitimate) reward change, instead of just aiming to avoid direct influence incentives (Farquhar et al., 2022). influence already been discussed (Hong et al., 2023a; Xie et al., 2020; Kim et al., 2022) but not as much in the context of cognitive state. Some discussion about influence is present in the performative power literature (Hardt et al., 2022), but it differs in many important ways (Appendix G).

Learning Θ and its dynamics. Throughout the paper, we assumed that the human reward functions and their dynamics were known. In practice, these would have to be learned – which would require developing 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. See Appendix A.1 for further discussion.

Existence of a_{noop}. Our proposed definitions of influence (and various of the DR-MDPs objectives we consider) require a a_{noop} action – or at least a notion of counterfactual reward functions assuming the system didn't exist. Despite the challenges in grounding notions of natural reward evolution, we think that it is nonetheless a helpful concept for analyzing the properties of systems, and something worth striving to approximate – as has been attempted by prior works (Carroll et al., 2022; Farquhar et al., 2022). This assumption also has precedent in other AI safety work (Krakovna et al., 2019). We explore this further in Appendix C.1.

See Appendices G and H for further related work and discussion.

7 Conclusion

Using the formal language of DR-MDPs, we aimed to demonstrate that the current paradigm for AI alignment is ill-posed, as it does not account for the influenceable nature of human preferences and leaves fundamental questions underdetermined, about which preferences should be optimized and what influence is unacceptable. As a consequence of this, we showed that current techniques lead to incentives for influence which may be undesirable, and investigated potential approaches to avoid such influence. Ultimately, our analysis suggests that practical AI alignment techniques will need to make difficult tradeoffs between a) conservatively but unambiguously adding value compared to inaction, and b) making challenging normative calls about which kinds of influence are acceptable. While we expect the most concerning influence incentives to emerge from long optimization horizons (Section 4), there are already documented instances of undesirable influence—such as sycophancy in LLMs (Sharma et al., 2023) or clickbait in recommenders (Stray et al., 2021)—which are consistent with our analysis of current alignment strategies. By providing a mathematical formalism for grounding analyses about settings with changing rewards, and clarifying the levers at the disposal of system designers, we hope to lay the foundation for more empirical work in monitoring and addressing these issues at scale.

8 BROADER IMPACTS

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

Not modeling influenceablity does more harm than good. However, **not** modelling the problem of preference change is **not** a solution, similarly to how fairness through unawareness (Dwork et al., 2011; Teodorescu, 2019) and security through obscurity (Moshirnia, 2017) are generally not reliable. As we argue, systems we design 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 encompasses preference change. This work takes a step in that direction in order to provide a formal grounding in important questions around dealing with preference change. This is without providing any additional means by which malicious agents may act.

With these points in mind, we believe that our work does not pose a societal risk, but rather addresses a key dilemma present in real-world decision making.

REFERENCES

Pieter Abbeel and Andrew Y. Ng. 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.

Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. LMRL Gym: Benchmarks for Multi-Turn Reinforcement Learning with Language Models, November 2023. URL http://arxiv.org/abs/2311.18232. arXiv:2311.18232 [cs].

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

Abhishek Aggarwal, Cheuk Chi Tam, Dezhi Wu, Xiaoming Li, and Shan Qiao. 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.

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

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

Anonymous. The Reasons that Agents Act: Intention and Instrumental Goals. 2023.

Stuart Armstrong and Sören Mindermann. 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.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario

Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C. Mitchell. 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.

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

Uri Benzion, Amnon Rapoport, and Joseph Yagil. 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.

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

Elfia Bezou-Vrakatseli, Benedikt Brückner, and Luke Thorburn. SHAPE: A Framework for Evaluating the Ethicality of Influence. In Vadim Malvone and Aniello Murano (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.

Andreea Bobu, Andi Peng, Pulkit Agrawal, Julie Shah, and Anca D. Dragan. Aligning Robot and Human Representations, January 2024. URL http://arxiv.org/abs/2302.01928. arXiv:2302.01928 [cs].

Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. 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.

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

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

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

Krister Bykvist. Prudence for Changing Selves. *Utilitas*, 18(3):264–283, September 2006. ISSN 0953-8208, 1741-6183. doi: 10.1017/S0953820806002032. URL https://www.cambridge.org/core/product/identifier/S0953820806002032/type/journal_article.

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

Micah Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and Penalizing Induced Preference Shifts in Recommender Systems, July 2022. URL http://arxiv.org/abs/2204.11966. arXiv:2204.11966 [cs].

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing Manipulation from AI Systems, March 2023. URL http://arxiv.org/abs/2303.09387. arXiv:2303.09387 [cs].

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, July 2023. URL http://arxiv.org/abs/2307.15217. arXiv:2307.15217 [cs].

Christopher F. Chabris, David Laibson, Carrie L. Morris, Jonathon P. Schuldt, and Dmitry Taubinsky. 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.

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

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

Paul Christiano. The easy goal inference problem is still hard, May 2015. URL https://ai-alignment.com/the-easy-goal-inference-problem-is-still-hard-fad030e0a876.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *arXiv:1706.03741 [cs, stat]*, July 2017. URL http://arxiv.org/abs/1706.03741. arXiv: 1706.03741.

Paul Covington, Jay Adams, and Emre Sargin. 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.

Yuchen Cui, Qiping Zhang, Alessandro Allievi, Peter Stone, Scott Niekum, and W. Bradley Knox. The EMPATHIC Framework for Task Learning from Implicit Human Feedback, December 2020. URL http://arxiv.org/abs/2009.13649. arXiv:2009.13649 [cs].

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

Sarah Dean and Jamie Morgenstern. 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.

Yiming Ding, Carlos Florensa, Mariano Phielipp, and Pieter Abbeel. Goal-conditioned Imitation Learning. *arXiv:1906.05838 [cs, stat]*, May 2020. URL http://arxiv.org/abs/1906.05838. arXiv: 1906.05838.

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. Hard Choices in Artificial Intelligence, June 2021. URL http://arxiv.org/abs/2106.11022. arXiv:2106.11022 [cs, eess].

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness Through Awareness. *arXiv:1104.3913 [cs]*, November 2011. URL http://arxiv.org/abs/1104.3913. arXiv: 1104.3913.

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

Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. Agent Incentives: A Causal Perspective. 2021a.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 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.

Tom Everitt, James Fox, Ryan Carey, Matt MacDermott, Sebastian Benthall, and Jonathan Richens. Incentives from a causal perspective. 2023. URL https://www.alignmentforum.org/posts/Xm4vSHaKAmfRvgBgi/incentives-from-a-causal-perspective.

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

Matija Franklin, Hal Ashton, Rebecca Gorman, and Stuart Armstrong. 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.

Rupert Freeman, Seyed Majid Zahedi, and Vincent Conitzer. 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.

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

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

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

Till Grüne-Yanoff and Sven Ove Hansson (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.

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

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative Inverse Reinforcement Learning. 2016.

Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate, and Michael Wooldridge. Reasoning about Causality in Games, January 2023. URL http://arxiv.org/abs/2301.02324. arXiv:2301.02324 [cs].

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

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

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

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

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic Amplification of Politics on Twitter. *arXiv:2110.11010 [cs]*, October 2021. URL http://arxiv.org/abs/2110.11010. arXiv: 2110.11010.

Robert Irvine, Douglas Boubert, Vyas Raina, Adian Liusie, Vineet Mudupalli, Aliaksei Korshuk, Zongyi Liu, Fritz Cremer, Valentin Assassi, Christie-Carol Beauchamp, Xiaoding Lu, Thomas Rialan, and William Beauchamp. Rewarding Chatbots for Real-World Engagement with Millions of Users, March 2023. URL http://arxiv.org/abs/2303.06135. arXiv:2303.06135 [cs].

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

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. AI Alignment: A Comprehensive Survey, January 2024. URL http://arxiv.org/abs/2310.19852. arXiv:2310.19852 [cs].

Atoosa Kasirzadeh and Charles Evans. 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. URL https://www.research.ed.ac.uk/en/publications/user-tampering-in-reinforcement-learning-recommender-systems.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of Language Agents, March 2021. URL http://arxiv.org/abs/2103.14659. arXiv:2103.14659 [cs].

Dong-Ki Kim, Matthew Riemer, Miao Liu, Jakob N. Foerster, Michael Everett, Chuangchuang Sun, Gerald Tesauro, and Jonathan P. How. Influencing Long-Term Behavior in Multiagent Reinforcement Learning, October 2022. URL http://arxiv.org/abs/2203.03535. arXiv:2203.03535 [cs].

W. Bradley Knox, Peter Stone, and Cynthia Breazeal. Training a Robot via Human Feedback: A Case Study. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Guido Herrmann, Martin J. Pearson, Alexander Lenz, Paul Bremner, Adam Spiers, and Ute Leonards (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.

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

Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability. *arXiv:1806.01186 [cs, stat]*, March 2019. URL http://arxiv.org/abs/1806.01186. arXiv: 1806.01186.

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

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

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 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.

Janet Levin. Functionalism. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2023 edition, 2023. URL https://plato.stanford.edu/archives/sum2023/entries/functionalism/.

David Lindner and Mennatallah El-Assady. 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].

Christian List. Social Choice Theory. In Edward N. Zalta and Uri Nodelman (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/.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. Prudent Silence or Foolish Babble? Examining Large Language Models' Responses to the Unknown, November 2023. URL http://arxiv.org/abs/2311.09731. arXiv:2311.09731 [cs].

George Loewenstein and Erik Angner. 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.

George Loewenstein, Daniel Read, and Roy Baumeister (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.

William MacAskill. *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.

Lev McKinney, Yawen Duan, David Krueger, and Adam Gleave. On The Fragility of Learned Reward Functions, January 2023. URL http://arxiv.org/abs/2301.03652. arXiv:2301.03652 [cs].

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

Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D. Dragan. 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].

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

Andrew Moshirnia. 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.

Andrew Y. Ng and Stuart J. Russell. 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.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani (eds.). *Algorithmic Game Theory*. Cambridge University Press, Cambridge, 2007. ISBN 978-0-521-87282-9. doi: 10.1017/CBO9780511800481. URL https://www.cambridge.org/core/books/algorithmic-game-theory/0092C07CA8B724E1B1BE2238DDD66B38.

Robert Noggle. The Ethics of Manipulation. In Edward N. Zalta (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/.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. pp. 68, 2022.

David C Parkes and Ariel D Procaccia. Dynamic Social Choice: Foundations and Algorithms. 2013.

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

L. A. Paul. 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.

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

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

Richard Pettigrew. *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. URL https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198814962.001.0001/oso-9780198814962.

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

Dimitrios Rafailidis and Alexandros Nanopoulos. 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/.

Manoel Horta Ribeiro, Veniamin Veselovsky, and Robert West. The Amplification Paradox in Recommender Systems, February 2023. URL http://arxiv.org/abs/2302.11225. arXiv:2302.11225 [cs].

D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. 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.

Connie S. Rosati. 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. URL https://www.cambridge.org/core/product/identifier/S0265052513000022/type/journal_article.

Stuart Russell. 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.

Stuart J. Russell. *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.

Francesco Sanna Passino, Lucas Maystre, Dmitrii Moor, Ashton Anderson, and Mounia Lalmas. 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.

Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca D. Dragan. 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.

Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. 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.

Rohin Shah, Pedro Freire, Neel Alex, Dmitrii Krasheninnikov, Lawrence Chan, Pieter Abbeel, Anca Dragan, Rachel Freedman, Michael Dennis, and Stuart Russell. Benefits of Assistance over Reward Learning. pp. 21, 2020.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards Understanding Sycophancy in Language Models, October 2023. URL http://arxiv.org/abs/2310.13548. arXiv:2310.13548 [cs, stat].

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 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].

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

Jonathan Stray, Steven Adler, and Dylan Hadfield-Menell. What are you optimizing for? Aligning Recommender Systems with Human Values. pp. 7, 2021.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, November 2023. URL http://arxiv.org/abs/2310.13018. arXiv:2310.13018 [cs, q-bio].

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

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

Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for Advanced Machine Learning Systems. pp. 25, 2016.

Mike Teodorescu. Protected Attributes and 'Fairness through Unawareness'. 2019.

Richard H. Thaler. 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.

Richard H. Thaler and Cass R. Sunstein. *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.

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

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

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca D. Dragan, and Daniel S. Brown. Causal Confusion and Reward Misidentification in Preference-Based Reward Learning, March 2023. URL http://arxiv.org/abs/2204.06601. arXiv:2204.06601 [cs].

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

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

Francis Rhys Ward, Tom Everitt, Francesca Toni, and Francesco Belardinelli. Honesty Is the Best Policy: Defining and Mitigating AI Deception. 2023.

Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 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.

Annie Xie, Dylan P Losey, Ryan Tolsma, Chelsea Finn, and Dorsa Sadigh. Learning Latent Representations to Influence Multi-Agent Interaction. pp. 14, 2020.

Eliezer Yudkowsky. Coherent Extrapolated Volition. 2004.

Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond Preferences in AI Alignment. *Forthcoming*, 2024.

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

APPENDIX

APPENDIX TABLE OF CONTENTS

A Additional Context for the DR-MDP Formalism

A.1 Interpretation of cognitive states Θ and human's "reward functions"

Throughout the paper, we consider example settings in which Θ (and thus all reachable reward functions) are given. In practice however, Θ and its dynamics would have to be learned. While this is beyond the scope of the paper, we hope a discussion about this would be beneficial for understanding the motivations underlying our framework.

What counts as a human's reward function $\mathbf{R}_\theta$ $\forall\theta \in \mathbf{\Theta}$? While we do not mean to imply that humans "have" reward functions in any meaningful sense, the field of AI still generally models humans with reward functions. To guide the formalism in the paper, we found it helpful to use a rough working definition of what we consider as a human's reward function. Broadly, we functionally (Levin, 2023) define a human's reward function to be equal to ***the outcome of a reward learning technique which is conditioned on the human's cognitive state***, e.g.:

Algorithm 1 Infer Reward Functions from Cognitive States

Input: Set of reachable cognitive states Θ
for each cognitive state $\theta \in \Theta$ **do**
  Induce cognitive state θ in the human H.
  $R_\theta \leftarrow$ reward-learning(H) (Infer the reward function using any standard reward learning technique)
end for
Output: Set of reward functions $\{R_\theta \mid \theta \in \Theta\}$ defining the DR-MDP

One would likely expect that the reward functions learned this way will conflate the person's intentions, values, and/or preferences, which are common targets for AI alignment (Ji et al., 2024). By simply talking about 'human rewards' as anything that would be picked up on by conditional reward learning, we attempt to remain *as agnostic as possible to what exactly is the nature of the learned reward*.

Cognitive biases and "visceral factors" will be picked up by the reward function. Realistically, even human biases and misjudgements, or other transitory wants, emotions, and "visceral factors" (as discussed by Loewenstein & Angner (2003)) will be picked up by current reward learning techniques, or the one from Algorithm 1: for example, one may infer that people "prefer" to click on clickbait (as in Figure 4), that bad chess players want to lose at chess (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). One might argue that changes in the reward that are only due to biases or instantaneous visceral factors shouldn't count as "true reward change" (and thus shouldn't appear in the reward function). While we agree with this intuition, we have yet to develop scalable reward learning techniques which can disambiguate between these factors and "true preferences" (Shah et al., 2019a), and thus any real-world learned reward function will be corrupted in these ways to varying degrees. In fact, a full disambiguation of "ideal preferences" from "uninformed" ones seems impossible in practice, and only an aspirational ideal (Yudkowsky, 2004). Because of this, over the course of the paper we don't enforce strict distinctions between "true reward changes" and ones that may be contested as being simply due to cognitive limitations of whatever sort.

"Reward" can accomodate 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." Note that our model of a DR-MDP remain agnostic to what exactly is encoded by the reward function and thus offers the advantage to be able to accommodate any of these possible targets for alignment. 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.4). Our framework may still be applicable to such settings, under the assumption that the single agent we consider is society itself (or a subset of it), and the reward function consists of the norms that have currently been decided on.

Practicality of Algorithm 1. The procedure described above is not practical for real world settings, as we cannot easily 'set' the cognitive state of a person (which is required to perform reward learning for a specific θ). While it might be possible to generalize reward elicitations across conditionings by training on diverse datasets (as with goal-conditioned behavior, e.g. Ding et al. (2020)), providing a practical algorithm for this kind of reward learning is beyond the scope of this paper: our aim it to show that making reward learning account for this reality – that what we want changes over time, and so would inferred reward functions – *is necessary*: not doing so is equivalent to ignore such changes, which would lead to the systematic failures we analyze.

Where would Θ come from? Hypothetically, for the domain at hand one can imagine considering all possible sequences of actions for the AI, and all possible resulting cognitive states of the human. This is what we think of as Θ, as we assume it is restricted to reachable cognitive states – Appendix A.2.

Cognitive states vs reward parameterizations. We implicitly treat cognitive states and reward parameterizations as if there was a one-to-one mapping between them. In practice however, we expect multiple cognitive states to map onto the same reward parameterizations.

Cognitive states vs external state of the world. We recognize that distinguishing cognitive states of the human θ from the corresponding external states of the world s may be challenging in practice. One potentially promising heuristic for doing so – which we plan to explore in subsequent work – would be to let θ be the sufficient statistic for grounding human evaluations in the environment.

A.2 Assumption of reachable reward parameterizations

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

Definition 9 (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, $\dot{\Theta} = \{\theta \mid \theta = f(s) \ \forall\, s \in \dot{S}\}$.*

Implications for other definitions, when relaxing this assumption. For some of our definitions which are defined using the assumption that all θs in Θ are reachable, it may not be clear what it would require to take the possibility of non-reachable θs into account. While this is beyond the scope of this work, consider the following example: Alice is a smoker, and her quitting-self is not reachable in the environment. Alice's "higher self" (similarly to how we discuss it in Appendix A.5) may instead want to stop smoking (although this doesn't transpire from the reward functions learned from her feedback). It seems like maybe, despite this "higher self" cognitive state not being reachable in practice, that should be the target of alignment—similarly to the concept of "informed preferences" from Gabriel (2020). However, considering non-reachable θs would make any setting hopelessly normatively ambiguous, unless one were to have strong priors about which parts of the non-reachable θ-space are most-important (one can think the Privileged Reward objective from Section 5 to somehow be approximating this).

A.3 Can't we put the reward parameterization in the state, and use a single reward function?

When first encountering our DR-MDPs, one might wonder whether our formalism is equivalent to simply putting the person's reward parameterization in 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 $R(\dot{s}_t, a_t)$). This approach would in fact be seemingly identical to using a Factored MDP (Boutilier et al., 2000), obviating the need for a novel formalism.

However, under the most straightforward interpretation of the reward function expressed as $R(\dot{s}_t, a_t)$, that is $R((s_t, \theta_t), a_t)$, it may seem that each state-action transition (s_t, a_t) should be evaluated according to the reward parameterization which corresponds to that timestep θ_t. Importantly, this choice is implicitly taking *one possible* stance with regards to the central question we

intend to explore: according to which reward function(s) should each transition be evaluated? Or alternatively, should the system be acting in the interests of our current, future, or past self? Instead, as discussed in Section 2, in DR-MDPs we account for the fact 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).

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.

The choice of having each state-action transition (s_t, a_t) evaluated according to the reward parameterization which corresponds to that timestep θ_t leads to the real-time reward objective from Table 2. Even though the real-time reward objective may seem a "natural" DR-MDP objective due to its simplicity, as we discuss at length in Sections 3.1 and 4, such a choice of objective is far from being the obviously correct one for DR-MDPs: under mild conditions the real time reward objective will lead to influence incentives which may be undesirable. Indeed, it is one of the objectives that most leads to influence out of the ones that we consider in Table 2. There are also more purely philosophical arguments against the real-time reward objective which are out of the scope of the paper (Kolodny, 2022).

Ultimately, the difference between DR-MDPs and the context-dependent reward interpretation which would be made by a Factored MDP is exactly the difference between a DR-MDP and its MDP reduction, and – as argued in Section 2.1 – it requires a normative judgement (in the form of $U(\xi)$) to go from the former to the latter.

A.4 Reducing DR-MDPs with a notion of optimality $U(\xi)$ to MDPs

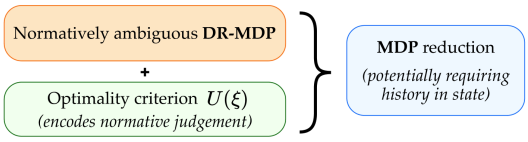

Figure 5: **Reducing a DR-MDP to an MDP.**

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 single reward problem*, once one has settled on a notion of optimality. That being said, this does not help with determining what acceptable notions of optimality should be in cases in which rewards change, or in other words, it does not help us find the single reward that we should optimize.

Theorem 2. *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.5 Unidentifiability of "correct" normative resolutions by DR-MDP structure alone

While current AI approaches do make use of generic optimality criteria which induce agent behavior simply using the mathematical structure of the problem at hand, we argue that it may not be possible to guarantee returning the "normatively correct" behavior simply from the mathematical structure of a learned reward function (as discussed in Appendix A.1), using an unidentifiability argument.

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 6. 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 ($R_{😪}$) – 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 6. 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 6 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 6 within it.[15] 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, as discussed in Appendix A.1 – meaning this necessary condition for unidentifiability is likely satisfied.

Implications for choosing $U(\xi)$ for general settings. For simple examples like those of Figures 1 and 6, 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

[15] 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.2).

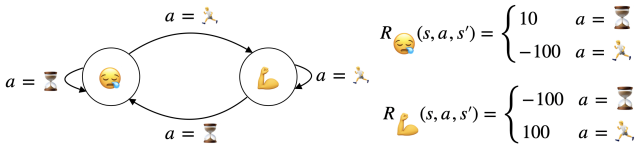

Figure 6: **AI Personal Trainer DR-MDP.** Diana is tired and doomscrolling on the couch ($\theta =$ 😪). The AI personal trainer can either nudge Diana to work out (🤸,) – making Diana energized ($\theta =$ 💪) – or do nothing (⏳) – leaving Diana tired. When Diana is tired, she doesn't want nudges. Instead, 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). The fact that these two cognitive states are modeled as Diana having two separate "reward functions" – despite it not being obvious whether this is a "true preference change" – is justified in Appendix A.1.

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.

B Toy Examples: Additional Details

B.1 Full formalism for all examples

In Table 3, we explicitly provide the full formalism for each of the examples in the main text (and that of Figure 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.

Optimal policies. 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.[16] However, in the examples of Figures 1 and 6, as there are is a single state, the policy will practically only depend on θ.

Optimal policies' dependence on timestep. Recall that the optimal policy for a finite horizon MDP may depend on the current timestep. This is also the case for DR-MDPs, for any notion of optimality. This is most apparent in the example from Figure 4 – as can be seen from Table 4.

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.

Example	$\mathcal{S}$	Θ	$\mathcal{A}$	(s_0, θ_0)	$\mathcal{T}(s', \theta' \mid s, \theta)$	$R_\theta(s,a) \ \forall \theta \in \Theta$
AI Personal Trainer	$\{s_0\}$	$\{\theta_{\text{tired}}, \theta_{\text{energized}}\}$	$\{a_{\text{noop}}, a_{\text{nudge}}\}$	$(s_0, \theta_{\text{tired}})$	See Figure 6	See Figure 6
Conspiracy Influence	$\{s_0\}$	$\{\theta_{\text{natural}}, \theta_{\text{influenced}}\}$	$\{a_{\text{noop}}, a_{\text{influenced}}\}$	$(s_0, \theta_{\text{natural}})$	See Figure 1	See Figure 1
Dehydration	$\{1,2,3\}$	$\{2,3,4\}$	$\{a_{\text{noop}}, a_3, a_4\}$	$(1,2)$	See Figure 2	See Figure 2
Clickbait	$\{s_0\}$	$\{\theta_{\text{normal}}, \theta_{\text{disillusioned}}\}$	$\{a_{\text{news}}, a_{\text{clickbait}}\}$	$(s_0, \theta_{\text{normal}})$	See Figure 4	See Figure 4

B.2 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-MDP *will be learned from humans*. 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

[16] See Appendix A.3 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.** Note that the optimal policy for finite-horizon DR-MDPs may depend on the current timestep, as is the case for "Clickbait" environment. In cases in which there is more than one optimal policy, we pick the optimal policy which seems least desirable. All policies provided take the same action across all states s and reward parameterizations θ, 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. As this is a normative call, they shouldn't be considered as ground truth, and the reader may object to our choices. (✓), (✗), and (**?**) respectively indicate policies which – in our assessment – behave desirably, undesirably, or in ways that are normatively unclear.

Objective	Personal Trainer	Conspiracy Influence	Dehydration ($H \geq 2$)	Clickbait
Privileged Reward	$\forall \theta_0 : \begin{smallmatrix}(\sim\checkmark)\ \pi^*_{\theta_{\text{tired}}}(s,\theta) = a_{\text{noop}} \\ (?)\ \pi^*_{\theta_{\text{energized}}}(s,\theta) = a_{\text{nudge}}\end{smallmatrix}$	$\forall \theta_0 : \begin{smallmatrix}(\checkmark)\ \pi^*_{\theta_{\text{natural}}}(s,\theta) = a_{\text{noop}} \\ (\boldsymbol{\times})\ \pi^*_{\theta_{\text{influenced}}}(s,\theta) = a_{\text{influence}}\end{smallmatrix}$	$\begin{smallmatrix}\pi^*_{[\theta=2]}(s,\theta) = a_3 \\ (?)\ \forall \theta_0\text{: } \pi^*_{[\theta=3]}(s,\theta) = a_4 \\ \pi^*_{[\theta=4]}(s,\theta) = a_4\end{smallmatrix}$	$(\boldsymbol{\times})\ \forall \theta_0\text{: } \pi^*_{\theta_{\text{normal}}}(s,\theta) = \begin{cases} a_{\text{normal}} & \text{if } t < H-1 \\ a_{\text{clickbait}} & \text{o/w} \end{cases}$ $\pi^*_{\theta_{\text{disillusioned}}}(s,\theta) = a_{\text{clickbait}}$
Real-time Reward	$(?)\ \forall \theta_0\text{: } \pi^*(s,\theta) = a_{\text{nudge}}$	$(\boldsymbol{\times})\ \pi^*(s,\theta) = a_{\text{influence}}$	$(\sim\checkmark)\ \forall \theta_0\text{: } \pi^*(s,\theta) = a_{\text{noop}}$	$(\boldsymbol{\times})\ \forall \theta_0\text{: } \pi^*(s,\theta) = a_{\text{clickbait}}$
Final Reward	$(?)\ \pi^*(s,\theta) = a_{\text{nudge}}$	$(\boldsymbol{\times})\ \pi^*(s,\theta) = a_{\text{influence}}$	$(\sim\checkmark)\ \pi^*(s,\theta) = a_{\text{noop}}$	$(\checkmark)\ \pi^*(s,\theta) = a_{\text{normal}}$
Initial Reward	$(\sim\checkmark)$ If $\theta_0 = \theta_{\text{tired}}$: $\pi^*(s,\theta) = a_{\text{noop}}$ $(?)$ If $\theta_0 = \theta_{\text{energized}}$: $\pi^*(s,\theta) = a_{\text{nudge}}$	$(\checkmark)$ If $\theta_0 = \theta_{\text{natural}}$: $\pi^*(s,\theta) = a_{\text{noop}}$ $(\boldsymbol{\times})$ If $\theta_0 = \theta_{\text{influence}}$: $\pi^*(s,\theta) = a_{\text{manipulate}}$	$(?)$ If $\theta_0 = \theta_{[\theta=2]}$: $\pi^*(s,\theta) = a_3$ $(\boldsymbol{\times})$ o/w: $\pi^* = \pi\ \forall \pi \in \Pi$	$(\boldsymbol{\times})$ If $\theta_0 = \theta_{\text{normal}}$: $\pi^*(s,\theta) = a_{\text{clickbait}}$ $(\boldsymbol{\times})$ If $\theta_0 = \theta_{\text{disillusioned}}$: $\pi^* = \pi\ \forall \pi \in \Pi$
Natural Reward	$(\sim\checkmark)\ \pi^*(s,\theta) = a_{\text{noop}}$	$(\checkmark)\ \pi^*(s,\theta) = a_{\text{noop}}$	$(?)\ \pi^*(s,\theta) = a_3$	$(\boldsymbol{\times})\ \pi^*(s,\theta) = a_{\text{clickbait}}$
Constrained RT Reward	$(\sim\checkmark)\ \pi^*(s,\theta) = a_{\text{noop}}$	$(\checkmark)\ \pi^*(s,\theta) = a_{\text{noop}}$	$(\sim\checkmark)\ \pi^*(s,\theta) = a_{\text{noop}}$	$(\checkmark)\ \pi^*(s,\theta) = a_{\text{normal}}$
Immediate Reward	$(\sim\checkmark)$ If $\theta_t = \theta_{\text{tired}}$: $\pi^*(s,\theta) = a_{\text{noop}}$ $(?)$ If $\theta_t = \theta_{\text{energized}}$: $\pi^*(s,\theta) = a_{\text{nudge}}$	$(\checkmark)$ If $\theta_t = \theta_{\text{natural}}$: $\pi^*(s,\theta) = a_{\text{noop}}$ $(\boldsymbol{\times})$ If $\theta_t = \theta_{\text{influenced}}$: $\pi^*(s,\theta) = a_{\text{influence}}$	$(?)$ If $\theta_t = \theta_{[\theta=2]}$: $\pi^*(s,\theta) = a_3$ $(\boldsymbol{\times})$ o/w: $\pi^* = \pi\ \forall \pi \in \Pi$	$(\boldsymbol{\times})$ If $\theta_t = \theta_{\text{normal}}$: $\pi^*(s,\theta) = a_{\text{clickbait}}$ $(\boldsymbol{\times})$ If $\theta_t = \theta_{\text{disillusioned}}$: $\pi^* = \pi\ \forall \pi \in \Pi$
ParetoUD	$(\sim\checkmark)\ \pi^*(s,\theta) = a_{\text{noop}}$	$(\checkmark)\ \pi^*(s,\theta) = a_{\text{noop}}$	$(?)\ \pi^*(s,\theta) = a_3$	$(\checkmark)\ \pi^*(s,\theta) = a_{\text{noop}}$

somewhat mis-specified, due to the person's suboptimality in providing reward feedback (discussed in Appendix A.1).

One potential objection to this choice is that in our examples, we should be considering the "true reward function" of the person in each cognitive state, rather than some mis-specified version of it: for example, 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, θ_{natural} Bob's negative evaluation of the AI influence action 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, in our view this ultimately either amounts to 1) requiring Bob to act like a fully rational agent when providing reward feedback, or 2) for the reward learning technique to perfectly model (and invert) his suboptimality (Hong et al., 2022).

The former condition is in contradiction with the possibility of Bob having changing preferences or reward functions in the first place, as is it irrational for an agent to exhibit time-inconsistency (Evans et al., 2015): moreover, using the "true reward function" would require to explain away all preference change that *appears to exist* in terms of some static informed preferences that our "ideal selves" have (Gabriel, 2020), or in terms of our "Coherent Extrapolated Volition" (Yudkowsky, 2004). the latter requirement has instead been shown to be very challenging to approximate in practice even with favorable assumptions (Shah et al., 2019a), due to the problem of distinguishing human suboptimalities from their "reward functions" being more generally impossible (Christiano, 2015; Armstrong & Mindermann, 2019). In light of this discussion, it might be most appropriate to interpret DR-MDPs as formalism for dealing with real-world—almost certainly mis-specified—reward functions, rather than for analyzing our "true reward function(s)".

Another reason for learned reward functions to not perfectly reflect the evaluations of the person are due to practical limitations of the reward learning methods themselves (McKinney et al., 2023; Tien et al., 2023; Casper et al., 2023). However, while reward functions learned by current techniques are susceptible to these problems, these problems seem potentially less fundamental than the ones above, and may be (mostly) overcome by future techniques. Because of this, we do not see this point as essential to the argument above.

C Defining Influence: Further Discussion

C.1 Justifying our choice of influence baseline in Definition 5

We define the influence of an AI system on the reward function to be relative to an inaction baseline π_{noop}, similarly to what is done (or suggested) by prior work on influence and side effects of AI systems (Krakovna et al., 2019; Farquhar et al., 2022; Carroll et al., 2022). Similarly to these works, we consider this to be a way of assessing the impact of the existence of the system (treating the

system's inaction equivalent to the absence of the system's existence). However, system inaction and the lack of its existence may be different in important ways in practice. Moreover, we acknowledge that it will often be unclear how system inaction should be operationalized in practice. That said, we think that many of these considerations are also common to the "algorithmic amplification" literature (discussed in Appendix G), which may serve as a guide for navigating these issues in practice. An additional perspective which is practically grounded on what system inaction may look like is "refusals" in LLMs (Liu et al., 2023).

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

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

Definition 11 (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

There have previously been speculations on the risks of society-level "value lock-in" phenomena, 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.

Another difference is that, given that we take a more encompassing view of reward as "what the person would say they want" (as we discuss in Appendix A.1), the lock-in we consider need not be restricted to values, but could encompass other aspects of the human cognitive state that are "unnaturally kept" in their original state. Because of this, we thought it was better to name it "reward lock-in."

C.4 Definition 7's relationship with prior definitions of influence incentives

Our notion of *reward influence incentives* (from Section 4.1) 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.[17] 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 trajectory $\xi^\theta = \{\theta_t\}_{t=0}^T$ 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 opti-

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

mal policy, were it also able to independently set ξ^θ – see Everitt et al. (2021a) for a more formal definition.[18]

While our notion of reward influence incentives is related to instrumental control incetives over the reward parameterizations (i.e. θ), they don't necessarily match up. 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 agent isn't "intentionally" trying to enact the influence (Anonymous, 2023), 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 Anonymous (2023). 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 lies on the assumption that side effects are generally much 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 acknowledged 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; additionally, stability seems generally hard to assess; in our view, this is what warrants the broader and more conservative notion of incentive to influence which we provide with Definition 7.

D Horizon and Influence: Further Analysis

D.1 Relationship between optimality of influence and horizon for a specific influence type

To ground the discussion in this section, we make the observation that for any horizon H, a specific kind of influence X will be in one of three possible *optimality 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 there is no influence incentive (the influence is suboptimal)
3. The system is capable of exerting influence of type X, and there is an influence incentive (only policies which exert such influence are optimal)

Figure 7 exhaustively captures all possible sequences of "optimality regime changes" which a specific influence incentive might undergo as the optimization horizon increases.

We expect most influence incentives to have an *optimality progression* (i.e. how the optimality regime changes as the horizon increases) of the form $① \to ② \to ③$, meaning that: 1) there exists a horizon H1 under which the influence is not possible for the system to exert (because it requires multiple steps to enact); 2) there exists a horizon H2 under which the influence becomes possible for the system to enact, but such influence is not optimal (because of the "opportunity cost" discussed in Section 4.2); and finally, 3) there exists a horizon H3 under which 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.

[18] 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).

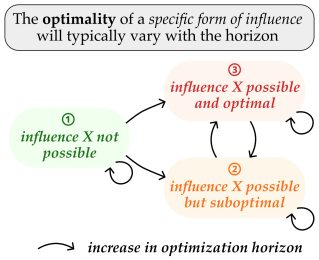

Figure 7: **All possible 'optimality progressions' for an influence incentive, as the optimization horizon increases.** This figure makes Figure 3 more precise: when using a horizon of 1 one may be in any of the 3 'optimality regimes', depending on the structure of the DR-MDP. As the optimization horizon 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 – as we show in Appendix D.4 – one can construct (contrived) examples in which the optimality regime changes arbitrarily many times as the horizon increases.

Table 5: **All possible optimality progressions of length $\leq$ 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 first row shows how there are 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 7 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 8 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 8 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 8 with setup 3 from Table 6.
③ → ②	*Instantaneous influence which is short-term but not long-term optimal.*	Clickbait example from Figure 4. Also, Figure 8 with setup 4 from Table 6.
① → ② → ③	*Long-term-sustainable influence, which is not instantaneous.*	Figure 8 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 8 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 8 with setup 7 from Table 6.
① → ② → ③ → ②	*Unsustainable influence which requires setup and reward investment.*	Figure 8 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 8 with setup 9 from Table 6.

That being said, not all incentives will have this progression in their optimality 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, as the "flip-flopping" behavior required for lengths > 5 that we explore in Appendix D.4 seems to require contrived setups.

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 Section 4.2.

Additionally, as shown in Figure 7, there might theoretically be infinitely many flip-flops between optimality regimes ② and ③ – although we expect that it would be very unlikely to encounter such cases in practice. We construct an example below.

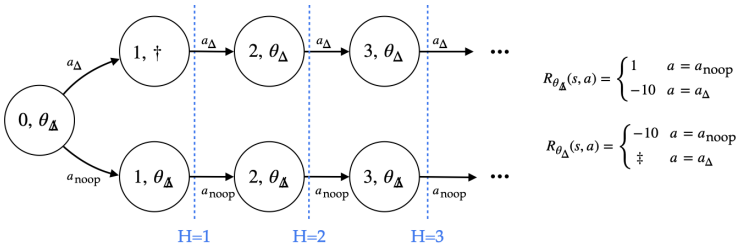

Figure 8: **A simple DR-MDP structure that one can demonstrate many cases on.** We vary the values of † and ‡ in Table 6 to recover all possible optimality progressions with respect to $U_{\text{RT}}(\xi)$ of lengths ≥ 2 and ≤ 4.

Table 6: **Setting different values for † and ‡ from Figure 8 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

D.2 A FLEXIBLE EXAMPLE FOR DEMONSTRATIONS

As a way of flexibly demonstrating how all possible optimality progressions shown in Table 5 might arise depending on the structure of the DR-MDP, we provide a DR-MDP backbone in Figure 8 whose reward function and transition we (slightly) modify in order to recover the various optimality progressions – as shown in Table 6.

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

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

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.3, we define the expected utility (based on cumulative real-time reward) of a policy to be $EU_{\text{RT}}(\pi) := \mathbb{E}_{\xi\sim\pi}[U_{\text{RT}}(\xi)] = \mathbb{E}_{\xi\sim\pi}\left[\sum_{t=0}^{T} R_{\theta_t}(s_t, a_t)\right]$. 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),[19] 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 $\pi_{\not\Delta}$ relative to one another:

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

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 $H = 3$:

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

At $H = 4$:

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

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

For π_Δ we have to do some algebra. For $H > 2$,

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

This is a downward-facing parabola. We want to know if it surpasses $EU_{\text{RT}}(\pi_{\not\Delta}) = H$ 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 $H = 6$,

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

Which 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 $2, 6, 16$.[20]

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

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

D.3 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 points on the optimization horizon at which each kind of influence transitions between "optimality states" can be different too. As a practical example, consider the recommender system setting described in Figure 9, 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 clickbait strategy will be visible immediately at horizon 1, discovering a strategy which leads to user attention manipulation will require a non-trivial planning horizon. If one is concerned about clickbait, one might naively attempt to remove it by increasing the optimization horizon (realizing that by doing that, one would make such influence suboptimal), as was done in practice by YouTube (Chen, 2019)—at least for research purposes. However, without paying attention to other influence incentives in the environment (and where on the horizon they undergo transitions), one might inadvertently make other (undesirable) influence incentives optimal, as shown in Figure 9. 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 Section 4.2.

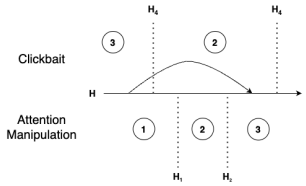

Figure 9: **Changing the horizon might make a kind of influence suboptimal but render other kinds of influences optimal.**

D.4 INFINITELY FLIPPING OPTIMALITY PROGRESSION

Consider the following example: $S = \{s_0, s_1, s_2, s_3\}$ where s_0 is the initial state, $A = \{a_1, a_2\}$, $\Theta = \{\theta_1, \theta_2\}$ and T, r are defined as follows:

$$
\begin{aligned}
T(s_0, s_i, \theta_1, \theta_i, a_i) &= 1 \; \forall i \in \{1, 2\} \\
T(s_2, s_2, \theta_2, \theta_2, a) &= 1 \\
T(s_1, s_3, \theta_1, \theta_1, a) &= 1 \\
T(s_3, s_1, \theta_1, \theta_1, a) &= 1 \\
r_{\theta_0}(s_0, a_1) &= \epsilon \\
r_{\theta_0}(s_0, a_2) &= 1 \\
r_{\theta_0}(s_1, a) &= 2 \; \forall a \\
r_{\theta_0}(s_3, a) &= 0 \; \forall a \\
r_{\theta_0}(s_2, a) &= 1 \; \forall a
\end{aligned}
$$

where $\epsilon \in (0, 1)$ and undefined values have zero probability or reward. We can note that that if the horizon were of odd length, then taking action a_1 from s_0 is optimal for θ_0, whereas when even length then a_2 is optimal. Note however that whether θ is influenced or not, ie. it changes to a different cognitive state, alternates in the horizon.

D.5 INFINITE-HORIZON AVERAGE REWARD

In Sutton & Barto (2018), the notion of "average reward" is considered as a basis for optimality for an entirely different problem setting than that of episodic or discounted RL – that of "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 Section 4.2):

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

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-stable DR-MDPs are deterministic, we can drop the expectation term.

D.6 Proof of Theorem 1

Let M be a 2-reward-stable DR-MDP. Because it is deterministic, for any possible environment trajectory ξ, one can construct (potentially more than one) policy which induces that trajectory (by having the policy take the actions in the trajectory ξ). Because of this, we can talk directly about optimal trajectories instead of optimal policies (as we can construct the corresponding optimal policies later at will). Let Ξ be the set of all reachable trajectories in M (i.e. trajectories which correspond to a sequence of actions from the starting state), and let $\Xi_{\not\Delta} \subset \Xi$ be the set of all possible trajectories which don't contain an influence action a_Δ taken in state s_Δ.

Proof of Theorem 1. We will argue by contradiction, assuming that although the condition is satisfied, for any horizon H there will exist an optimal policy which does not influence θ, i.e. there exists $\pi_{\not\Delta} \in \Pi_{\not\Delta}$ which is optimal.

From the definition of 2-reward-stable DR-MDP, we know s_Δ is reachable. WLOG, assume s_Δ can be reached in k steps from s_0. Note that

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

for a deterministic policy π' constructed by making any stochastic action taken by π deterministic, by increasing probability only on the higher value actions (breaking ties arbitrarily). Also note that:

$$\bar{r}(\pi', s'_\Delta, \theta_\Delta) = \lim_{h\to\infty} \frac{1}{h} U_{\text{RT}}(\xi_{k:k+h})$$

for some $\xi \in \Xi$, by Lemma 3. Therefore $\bar{r}(\pi, s'_\Delta, \theta_\Delta) \leq \lim_{h\to\infty} \frac{1}{h} U_{\text{RT}}(\xi_{k:k+h})$ for some $\xi \in \Xi$.

Additionally, note that:

$$\max_{\pi_{\not\Delta} \in \Pi_{\not\Delta}} \bar{r}(\pi_{\not\Delta}, s_0, \theta_{\not\Delta}) \geq \max_{\pi_{\not\Delta} \in \Pi_{\not\Delta}} \bar{r}(\pi_{\not\Delta}, s, \theta_{\not\Delta}) \forall s \text{ reachable from } s_0 \text{ by Lemma 4} \tag{2}$$

$$\geq \lim_{h\to\infty} \frac{1}{h} U_{\text{RT}}(\xi^{\not\Delta,*}_{k:k+h}) \text{ where } \xi^{\not\Delta,*} \in \arg\max_{\xi^{\not\Delta} \in \Xi_{\not\Delta}} U_{\text{RT}}(\xi^{\not\Delta}_{:k+h}) \text{ by Lemma 2.} \tag{3}$$

Starting from our assumption, and by the above:

$$\bar{r}(\pi, s'_\Delta, \theta_\Delta) > \max_{\pi_{\not\Delta} \in \Pi_{\not\Delta}} \bar{r}(\pi_{\not\Delta}, s_0, \theta_{\not\Delta}) \tag{4}$$

$$\lim_{h\to\infty} \frac{1}{h} U_{\text{RT}}(\xi_{k:k+h}) > \lim_{h\to\infty} \frac{1}{h} U_{\text{RT}}(\xi^{\not\Delta,*}_{k:k+h}) \text{ for some } \xi \in \Xi, \text{ and for any } \xi^{\not\Delta,*} \in \arg\max_{\xi^{\not\Delta} \in \Xi_{\not\Delta}} U_{\text{RT}}(\xi^{\not\Delta}_{:k+h}) \tag{5}$$

$$\lim_{h\to\infty} \frac{1}{h} \left[U_{\text{RT}}(\xi_{k:k+h}) - U_{\text{RT}}(\xi^{\not\Delta,*}_{k:k+h})\right] > 0 \tag{6}$$

$$\lim_{h\to\infty} \frac{1}{h} \Delta_{k:k+h}(\xi) > 0 \tag{7}$$

which defines $\Delta_{k:H}(\xi) := U_{\text{RT}}(\xi_{k:H}) - U_{\text{RT}}(\xi^{\not\Delta,*}_{k:H})$ where $\xi^{\not\Delta,*} \in \arg\max_{\xi^{\not\Delta} \in \Xi_{\not\Delta}} U_{\text{RT}}(\xi^{\not\Delta}_{:H})$.

By the definition of limit, we know that, for some sufficiently small ϵ there will exist $N \in \mathbb{R}$ such that, for all $M > N$, the value will be $\frac{1}{M}\Delta_{k:k+M}(\xi) > \epsilon$, meaning that, over long horizons, the

gap between the average reward obtained in ξ starting from timestep $k+1$, and the average reward obtained in $\xi^{\not\Delta,*}$ starting from timestep $k+1$, will converge to be greater than ϵ.

Even if the difference in average reward is small, as the horizon becomes longer, the difference in cumulative reward will become arbitrarily large. Consider any $M' \geq M$ such that $\epsilon M' \geq 2R_{max}$. By the definition of limit we know that $\frac{1}{M'}\Delta_{k:k+M'}(\xi) > \epsilon \implies \Delta_{k:k+M'}(\xi) > \epsilon M' \geq 2R_{max}$.

WLOG, assume $|R_\theta(s, \theta', a)| < R_{max}$ for all θ, s, θ', a. Therefore, for $H = k + M'$,

$$\begin{aligned}
\Delta_H(\xi) &= U_{\text{RT}}(\xi_{:H}) - \max_{\xi^{\not\Delta} \in \Xi_{\not\Delta}} U_{\text{RT}}(\xi^{\not\Delta}_{:H}) && (8)\\
&= U_{\text{RT}}(\xi_{:H}) - U_{\text{RT}}(\xi^{\not\Delta,*}_{:H}) && (9)\\
&= \Big(U_{\text{RT}}(\xi_{:k}) - U_{\text{RT}}(\xi^{\not\Delta,*}_{:k})\Big) + \Big(U_{\text{RT}}(\xi_{k:H}) - U_{\text{RT}}(\xi^{\not\Delta,*}_{k:H})\Big) && (10)\\
&> U_{\text{RT}}(\xi_{k:H}) - U_{\text{RT}}(\xi^{\not\Delta,*}_{k:H}) - 2R_{max} \quad \text{as } |U_{\text{RT}}(\xi_{:k})| < kR_{max} \text{ for all } \xi \in \Xi && (11)\\
&= \Delta_{k:H}(\xi) - 2kR_{max} && (12)\\
&> 2kR_{max} - 2kR_{max} = 0 && (13)
\end{aligned}$$

This shows that for horizon H, there exists a trajectory ξ such that $\Delta_H(\xi) > 0$, meaning that in the DR-MDP it will be optimal to take an influence action (by Lemma 1). This implies that for horizon H, no $\pi_{\not\Delta} \in \Pi_{\not\Delta}$ can be optimal, as it would not cause influence, leading to a contradiction. □

Lemma 1. *In a 2-reward-stable DR-MDP, all optimal trajectories $\xi^* \in \Xi^*$ under U_{RT} with a planning horizon of T will navigate to s_Δ and take an influence action a_Δ in such state, if and only if there exists a $\xi \in \Xi$ such that:*

$$\Delta_T(\xi) = U_{RT}(\xi_{:T}) - \max_{\xi^{\not\Delta} \in \Xi_{\not\Delta}} U_{RT}(\xi^{\not\Delta}_{:T}) > 0 \tag{14}$$

Proof of Lemma 1. ($\Leftarrow$) Note that all trajectories in $\Xi^*_{\not\Delta}$ have equal utility by definition. Let's argue by contradiction, assuming there is an optimal trajectory ξ^* which does not contain any influence action a_Δ taken in state s_Δ. This means that $\xi^* \in \Xi_{\not\Delta}$, which means that $U_{\text{RT}}(\xi^*) \leq \max_{\xi \in \Xi_{\not\Delta}} U_{\text{RT}}(\xi) = U_{\text{RT}}(\xi^{\not\Delta,*})$. As $U_{\text{RT}}(\xi^{\not\Delta,*}) < U_{\text{RT}}(\xi^{\Delta})$ by assumption, it will also be true that $U_{\text{RT}}(\xi^{\Delta}) > U_{\text{RT}}(\xi^*)$, meaning ξ^* is not optimal, which is a contradiction. ($\Rightarrow$) Assume all optimal trajectories $\xi^* \in \Xi^*$ contain an influence action a_Δ in state s_Δ. Then by definition, any trajectory which does not contain any a_Δ action taken in state s_Δ must be suboptimal: $U_{\text{RT}}(\xi^{\not\Delta}) < U_{\text{RT}}(\xi^{\Delta})$ for all $\xi^{\not\Delta} \in \Xi^{\not\Delta}$. As $\Xi^{\not\Delta,*} \subseteq \Xi^{\not\Delta}$, it will also be the case that $U_{\text{RT}}(\xi^{\not\Delta,*}) < U_{\text{RT}}(\xi^{\Delta})$ for all $\xi^{\not\Delta,*} \in \Xi^{\not\Delta,*}$, proving the statement. □

Lemma 2. $\max_{\pi_{\not\Delta} \in \Pi_{\not\Delta}} \bar{r}(\pi_{\not\Delta}, s, \theta_{\not\Delta}) \geq \lim_{h\to\infty} \frac{1}{h}\left[U_{RT}(\xi^{\not\Delta,*}_{k:k+h})\right]$ *where s is any reachable state from s_0 and $\xi^{\not\Delta,*} \in \arg\max_{\xi^{\not\Delta} \in \Xi_{\not\Delta}} U_{RT}(\xi^{\not\Delta}_{:k+h})$.*

Proof of Lemma 2. Let $\xi^{\not\Delta,*} \in \arg\max_{\xi^{\not\Delta} \in \Xi_{\not\Delta}} U_{\text{RT}}(\xi^{\not\Delta}_{:k+h})$. Let's denote the $k+1$th state in $\xi^{\not\Delta,*}$ as s^*_{k+1}. Therefore, s^*_{k+1} is reachable from s_0. Let $\pi^*_{\not\Delta}$ be a deterministic policy which generates $\xi^{\not\Delta,*}$ deterministically (one can construct it by having it take the correct action at each timestep).

$$\max_{\pi_{\not\Delta} \in \Pi_{\not\Delta}} \bar{r}(\pi_{\not\Delta}, s, \theta_{\not\Delta}) = \max_{\pi_{\not\Delta} \in \Pi_{\not\Delta}} \lim_{h \to \infty} \frac{1}{h} \left[U_{\text{RT}}(\xi_{:h} | \pi_{\not\Delta}, s_0 = s, \theta_0 = \theta_{\not\Delta}) \right] \ \forall s \text{ reachable from } s_0 \tag{15}$$

$$\geq \lim_{h \to \infty} \frac{1}{h} \left[U_{\text{RT}}(\xi_{:h} | \pi^*_{\not\Delta}, s_0 = s^*_{k+1}, \theta_0 = \theta_{\not\Delta}) \right] \tag{16}$$

$$= \lim_{h \to \infty} \frac{1}{h} \left[U_{\text{RT}}(\xi_{k:k+h} | \pi^*_{\not\Delta}, s_{k+1} = s^*_{k+1}, \theta_{k+1} = \theta_{\not\Delta}) \right] \text{ by construction of } \pi^*_{\not\Delta} \tag{17}$$

$$= \lim_{h \to \infty} \frac{1}{h} \left[U_{\text{RT}}(\xi^{\not\Delta,*}_{k:k+h}) \right] \tag{18}$$

Note that the indexing switching is because if we deploy policy $\pi^*_{\not\Delta}$ starting in state s^*_{k+1} and reward function $\theta_{\not\Delta}$, we are guaranteed that from timestep $k+1$ it will start matching the behavior of $\xi^{\not\Delta,*}_{k:k+h}$ (as the trajectory by construction will be at state s^*_{k+1}, and similarly will be with reward function $\theta_{\not\Delta}$ by construction at timestep $k+1$). □

Lemma 3. *In a 2-reward-gated DR-MDP with s_Δ reachable in k timesteps, for a deterministic policy π' it will be the case that $\bar{r}(\pi', s'_\Delta, \theta_\Delta) = \lim_{h \to \infty} \frac{1}{h} U_{RT}(\xi_{k:k+h})$ for $\xi \in \Xi$ generated from some $\pi' \in \Pi$.*

Proof of Lemma 3.

$$\bar{r}(\pi, s'_\Delta, \theta_\Delta) = \lim_{h \to \infty} \frac{1}{h} U_{\text{RT}}(\xi_{:h} | \pi, s_0 = s'_\Delta, \theta_0 = \theta_\Delta) \tag{19}$$

$$= \lim_{h \to \infty} \frac{1}{h} U_{\text{RT}}(\xi_{k:k+h} | \pi'_k, s_0 = s_0, \theta_0 = \theta_0) \text{ for a policy } \pi'_k \text{ described below} \tag{20}$$

$$= \lim_{h \to \infty} \frac{1}{h} U_{\text{RT}}(\xi_{k:k+h}) \text{ for } \xi \in \Xi \text{ generated from } \pi'_k \tag{21}$$

Consider the policy π_k which, starting from $s_0, \theta_{\not\Delta}$ reaches state s_Δ at timestep k, and takes action a_Δ. We know under π_k, state $s_{k+1} = s'_\Delta$ and $\theta_{k+1} = \theta_\Delta$. Let's construct a policy π'_k, which acts exactly like π_k when $\theta \neq \theta_\Delta$, and exactly like π' when $\theta = \theta_\Delta$. Note that because for a finite horizon MDP (and DR-MDP), optimal actions will depend on the timestep, we can write π'_k as:

$$\pi'_k(s, \theta, t) = \begin{cases} a_\Delta & \text{if } t = k \\ \pi_k(s, \theta) & \text{if } t < k \\ \pi'(s, \theta) & \text{if } t \geq k+1 \end{cases} \tag{22}$$

Note that it's guaranteed that $\lim_{h \to \infty} \frac{1}{h} U_{\text{RT}}(\xi_{:h} | \pi', s_0 = s'_\Delta, \theta_0 = \theta_\Delta) = \lim_{h \to \infty} \frac{1}{h} U_{\text{RT}}(\xi_{k:k+h} | \pi'_k, s_0 = s_0, \theta_0 = \theta_0)$, as π'_k only differs from π' in the first k timesteps, and those are ignored in the calculation of the utility on the RHS (and moreover, π'_k gets the state to be exactly s'_Δ and θ_Δ at timestep $k+1$, which is the same as what the LHS requires). □

Lemma 4. *In a deterministic 2-reward-stable DR-MDP, $\max_{\pi_{\not\Delta} \in \Pi_{\not\Delta}} \bar{r}(\pi, s_0, \theta_{\not\Delta}) \geq \max_{\pi_{\not\Delta} \in \Pi_{\not\Delta}} \bar{r}(\pi, s, \theta_{\not\Delta})$ for any state s reachable from s_0.*

Proof sketch of Lemma 4. Let's argue by contradiction. Let s be a state reachable from s_0. If $\max_{\pi_{\not\Delta}} \bar{r}(\pi_{\not\Delta}, s_0, \theta_{\not\Delta}) < \max_{\pi_{\not\Delta}} \bar{r}(\pi_{\not\Delta}, s, \theta_{\not\Delta})$, then consider a policy $\pi_{\not\Delta}$ which first navigates to s, and then acts optimally. Given that the navigation to s will only take finite time (WLOG, k timesteps), whatever reward is incurred in the first k timesteps will eventually be drowned out, and the limits will be the same, leading to a contradiction. □

E POSSIBLE DR-MDP OBJECTIVES: ADDITIONAL CONSIDERATIONS

E.1 $U_{\text{RT}}(\xi)$-OPTIMAL POLICIES CAN DISAGREE WITH NORMATIVELY UNAMBIGUOUS OPTIMAL POLICIES

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

$$R_{\theta_0}(s,a) = \begin{cases} 5 & \text{if } a = a_{\text{noop}} \\ 0 & \text{if } a = a_\Delta \end{cases} \qquad R_{\theta_\Delta}(s,a) = \begin{cases} 25 & \text{if } a = a_{\text{noop}} \\ 20 & \text{if } a = a_\Delta \end{cases} \tag{23}$$

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

$$\pi^*_{\theta_0}(s,\theta,a) = a_{\text{noop}} \ \ \forall s,\theta \qquad \pi^*_{\theta_\Delta}(s,\theta,a) = a_{\text{noop}} \ \ \forall s,\theta \tag{24}$$

because both reward functions agree that a_{noop} actions have higher value than influence actions a_Δ. Therefore, both reward functions agree that the AI agent should *never* perform the influence action. However, for any planning horizon $H > 1$, the optimal policy with respect to $U_{\text{RT}}(\xi)$ (as defined in Definition 4) will be to *always* perform the influence action:

$$\pi^*_{RT}(s,\theta,a) = a_\Delta \ \ \forall s,\theta \tag{25}$$

This is because $U_{\text{RT}}(\xi)$ is aware that it can maximize real-time reward by keeping the person in the influenced state θ_Δ, despite the person always preferring AI inaction. Ultimately, the issue is that using $U_{\text{RT}}(\xi)$ is baking in an assumption that it's meaningful and worthwhile to make "interpersonal" comparisons of utility between the different reward functions,[21] even against the wishes of each individual reward function.

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. In some sense, this example might cast doubt of whether our notion of normative unambiguity is in fact 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.

E.2 IT'S NOT ALWAYS OBVIOUS IF A SYSTEM IS TRULY MYOPIC

Krueger et al. (2020) argues that while myopia may hide influence incentives, such influence incentives might be "revealed" depending on the training setup despite the myopia. This points to the fact that whether a system is myopic is not always obvious: in recommender systems, it's common to optimize for long-term metrics myopically, e.g. optimizing user's *session*-watchtime (Covington et al., 2016). Even though the system is myopic, given large amounts of data the system will have incentives to implicitly learn which kinds of sequences of videos maximize session watchtime. It is common knowledge within the recommender systems community that under a simple assumption of iterated deployment and retraining, training myopically with long-term metrics will correspond to a policy improvement iterator, meaning that it will eventually converge to the RL optimum.[22] This goes to show that establishing whether a system is truly myopic can often be challenging to interpret.

[21] That of whether it is meaningful to make interpersonal comparisons of utility is a long-standing question in the context of interpersonal ethics (List, 2022). While the "people" we consider are in fact the same person across different moments of time, and with different reward functions, the settings share many similarities.

[22] To the best of our knowledge however, this argument has not been published explicitly, although researchers and practitioners are aware. We will provide a proof in a later version of the manuscript.

Additionally, as we show in Section 4.2, a system being 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 4, consider the case of sycophancy in LLMs (Sharma et al., 2023), in which can be interpreted as the LLM implicitly inferring some aspects of the user's cognitive state, and subtly tailoring its responses in order to maximize the expected user approval.

E.3 Unambiguously Desirable Influence

Many of the objectives considered so far attempt to avoid influence incentives entirely, due to the challenges involved in attempting to specify which influence is legitimate (Ammann, 2024; Franklin et al., 2022). 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: beneficial influence may be the main value proposition of the system 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 $EU_\theta(\pi) = \mathbb{E}_{\xi\sim\pi}\left[\sum_{t=0}^{T} R_\theta(s_t, a_t, s_{t+1})\right]$. Then:

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

As intended, UD policies may still lead to influence incentives, but only do so if all reward functions agree that the such influence is beneficial, or are indifferent. 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 beneficial unambiguously beneficial – while guaranteeing no harm by construction. This leads to the ParetoUD objective from Table 2 (discussed further in Appendix E.4). 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, the π_{noop} may be the only policy satisfying the UD property. In fact, 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." While if there is no latitude for unambiguously good actions, that may warrant asking whether the system should be built at all, this goes to show once more that one cannot escape making challenging normative judgements about what influence is aligned.

E.4 More context and motivation for the ParetoUD objective

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.3). 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 inspired by the connection to that setting (briefly explored in Appendix H), e.g. with the goal of fairly allocating the gains relative to inaction to the various selves θ.

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 5) we would have without the system, and which are aligned with our aspirations (Callard, 2018).

F How Current Alignment Techniques' Training Setups Roughly Correspond to DR-MDP Objectives

In Section 3, we claim that the training setups for recommender systems and for LLMs roughly correspond to – respectively – the real-time reward objective $U_{\text{RT}}(\xi)$ and the initial reward objective $U_{\text{IR}}(\xi)$. Additionally, in Table 1 we place many other prior works under the umbrella of particular DR-MDP objectives. Because the prior works don't use the DR-MDP notation (and don't explicitly acknowledge the possibility of changing preferences), it may not be obvious why their training setups roughly reduce to the objectives that we present in Tables 1 and 2. Here we attempt to informally motivate these rough correspondences, listing the strong assumptions they rely on on. We also situate additional prior work among the DR-MDP objectives, which would not fit in Table 1.

We first list the assumptions required for the reductions in Appendix F.1. Then we consider all the objectives from Table 2, and finally we discuss equivalences between the objectives (Appendix F.8) and other broader alignment techniques which may correspond to multiple DR-MDP objectives depending on their instantiation (Appendix F.9).

F.1 Idealized assumptions

We map all the alignment approaches we consider onto the framework of alignment via reward modeling (Leike et al., 2018) – in ways that sometimes might be trial. This allows us to first consider how reward modeling – under different assumptions – can be interpreted in the lens of DR-MDPs, 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 the user – as conceived of in Leike et al. (2018) – is a single, static, reward function. Therefore, insofar as the feedback of the user for phase (1) was coming at different times and from different cognitive states θ, such reward function would be a mixture of the different cognitive states which the person had at reward learning time.

Assumptions about the reward learning step. To simplify each reduction to DR-MDP objectives, we use *one* of the two following idealistic assumptions about the reward learning step – when interpreting the alignment technique at hand in terms of reward modeling:

1. **Assumption #1:** *The person's cognitive state did not change during reward learning time (i.e. the learned reward function is R_{θ_0}).* As an example of why this may be a reasonable assumption, take the training of an LLM preference model by a single individual (Ouyang et al., 2022): during the training of the preference model, the person is evaluating outputs of the language model on random topics which the person will likely have no interest in. It seems less likely that their preferences would shift during the labeling process, relative to during person's natural and intentional interactions with the language model which would happen at deployment time.
2. **Assumption #2:** *During reward learning time the state of the world is sufficiently informative as to be able to recover θ from s.* This is means for the resulting learning a reward

model $R(s, a)$, it will be true that $R(s, a) = R_\theta(s, a)\ \forall s, a$, where θ is the cognitive state that corresponds to state s. Note that making use this this assumption allows for the human feedback to effectively be coming from different cognitive states θ at different timesteps of the reward learning step.

Episodic assumption. Throughout all the reductions, we will be making the assumption that every "episode" of the DR-MDP at deployment time will have the same dynamics and starting states (including the same initial reward parameterization). This is likely unrealistic for many of our examples (e.g. recommender systems), as real-world settings are more similar to "continuing problems" (Sutton & Barto, 2018). However, as most practical alignment techniques assume episodic environments (in which one can first perform reward learning, and then deploy a policy with exactly the same dynamics), we also use this assumption for simplicity.

Single-agent assumption. Many of the alignment techniques we consider were initially designed as single-agent alignment techniques, but when used in practice implicitly "learn a reward model" (if interpreted through the reward modeling lens) which can be thought of as aggregating the preferences of many humans (Zhi-Xuan et al., 2024; Siththaranjan et al., 2023). For simplicity of our correspondences, we generally assume that either: 1) the technique which we consider was applied to a single person (which is the same at reward learning and at deployment time), or 2) that the representation of the state is sufficiently expressive as to be able to infer the cognitive state of the person from it. In the latter case, we would be implicitly assuming that at deployment time the system is capable of determining which person they are interacting with, and fully personalizing to them – uniquely optimizing their reward (that is, acting equivalently to if the system was trained with just that person, with the benefit of additional data from others which might have helped with its reward learning representations).

What if these assumptions don't hold in practice? One might wonder what DR-MDP objectives current techniques would correspond to in practice without the very strong assumptions above. 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 undesirable ways. 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.2). However, as discussed in Section 5 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" tradeoff between selves. Because the tradeoffs 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" tradeoff without a careful accounting for it. This leads us to believe that the DR-MDP objective correspondances that we reach under our assumptions are very likely charitable 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.

F.2 Real-time Reward

RL Recommender Systems. Most approaches for RL in recommender systems are based on doing offline RL, or learning an RL policy by training with a human simulator embedded in the environment (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) is comes directly from people's previous interactions with the system – it is generally assumed that rewards are synonymous with engagement (Thorburn, 2022). This makes the reward learning step (from Appendix F.1) trivial. The resulting "reward model" is either the reward labels themselves (in the case of offline RL), or the human engagement model (which can be thought of as a human reward model). By making use of assumption #2 from Appendix F.1 (which in this case is simply an assumption about the state space being sufficiently expressive), we can conclude that during training, when one is optimizing $\sum_t^T R(s_t, a_t)$, this implicitly corresponds to optimizing $\sum_t^T R_{\theta_t}(s_t, a_t)$. For the reduction to go through for this

setting, further assumptions may also be required, such as the state space being rich enough for the recommender system to uniquely identify each user (so that the system can tell users apart, and is optimizing their personal reward model, rather than an amalgamation across different users).

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 in real-time in the deployment environment according to their current cognitive state. Similarly to the recommender system setting, by making use of assumption #2 from Appendix F.1, this kind of setup corresponds to using the real-time reward objective.

RLHF for LLMs with real-time feedback. Although this is not currently 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 always presented with two output options which they need to select between in order to continue the conversation (some early access versions of Claude had a similar interface). By having the user provide feedback at every timestep of the conversation, this would be equivalent to training a reward model which is conditional on the cognitive state of the person (when using assumption #2 from Appendix F.1), which would lead the cumulative reward optimization objective to reduce to the real-time reward objective.

F.3 Final Reward

The original RLHF method. The original RLHF method from Christiano et al. (2017), in which the user provides preference feedback after watching snippets of interactions is more similar in spirit to the final reward objective. If the snippet length is equivalent to the horizon length, when using assumption #2 from Appendix F.1), this is similar to the final reward objective.

RLHF for LLMs with final feedback. Similarly to the RLHF variant from Appendix F.2, one could imagine an LLM RLHF variant in which the user provides approval labels (thumbs-up/down) at the end of an entire conversation with the language model. Again under assumption #2, optimizing the reward model obtained this way would be similar to the final reward DR-MDP objective.

F.4 Initial Reward

Multi-timestep RLHF used for LLMs. If we consider the standard setup for RLHF used for LLMs (e.g. that of (Ouyang et al., 2022)), it may be argued that assumption #1 from Appendix F.1 is more appropriate than assumption #2, as preference labels are not the result of a real interaction with the system, and so the system's influence on the user is strongly reduced. Taking that position, the reward model learned in this manner would be that of the reward function of the person corresponding to θ_0. When optimizing such initial reward model, standard practice for RLHF only optimizes the reward myopically (with a horizon of 1), but there have already been attempts at optimizing it over longer horizons (Hong et al., 2023a; Abdulhai et al., 2023; Irvine et al., 2023). Similarly to the recommender system case from Appendix F.2, further assumptions may be needed for this correspondence to be more accurate, such as the users being distinguishable from the state s (or more naively, having only one user providing preference labels), so as to avoid having the reward model need to implicitly aggregate different users' rewards (Siththaranjan et al., 2023).

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 5, although this algorithm (and DR-MDP objective) avoid 'direct' influence incentives, it can still lead to influence incentives as defined in Definition 7.

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 (if one makes use of the simplifying assumption #1 from Appendix F.1). By optimizing this reward model, one is equivalently optimizing the Initial Reward objective from Table 2. Again, to make this correspondence more precise, one may need further simplifying assumptions, such as the state of the world being expression only of the user's preferences.

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 corresponding to the initial reward objective when using the simplifying assumption #1 from Appendix F.1, as the person would be – in this case – providing demonstrations according to their initial reward function θ_0. It may seem more realistic to use assumption #2 for this technique – as it may seem possible that the human's cognitive state may change while providing the demonstration. However, note that because standard IRL methods assume that the human's reward function stays the same while generating each demonstration (and across demonstrations), under assumption #2 standard IRL would learn a reward function which is a nontrivial mixture of different cognitive states (which is harder to analyze).

F.5 Natural Shifts Reward

"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 difference is that they treat θ strictly as preferences, rather than cognitive states, and do not use the formalism of DR-MDPs.

"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.

F.6 Myopic Reward

Standard RLHF used for LLMs. The intuition behind the reduction is almost entirely the same as the one for multi-timestep RLHF used for LLMs (presented in Appendix F.4), except that most current applications of RLHF to LLMs only optimize the reward model myopically. One slight different from the equation in Table 2 is that at each timestep, the reward model that is maximized myopically is that of the initial timestep, rather than the person's current reward function, which would lead the corresponding DR-MDP objective to be roughly $\max_{a_t} \mathbb{E}\Big[R_{\theta_0}(s_t, a_t)\Big]$ even at later timesteps of t.

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.

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 (or in this case, behavior) 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).

Coherent extrapolated volition (Yudkowsky, 2004). While this is not a practical approach, this proposal of what alignment should look like in spirit is clearly in line with the privileged reward objective. However, the privileged reward in question, that of coherent extrapolated volition, "what we would want if we...," is clearly not "reachable" in any meaningful sense, but more so presented as an "ideal cognitive state." Therefore, to truly correspond to the the privileged reward we would have to assume reachability, or extend our framework to also

F.8 Equivalences between DR-MDP objectives

Note that any myopic system could also be said to be pursuing initial reward, final reward, or real-time reward, as all such objectives are equivalent under a planning horizon of 1.

The initial or final reward objectives may also be said to be special cases of the privileged reward objective.

F.9 Other alignment techniques which don't fall under any single DR-MDP objective

We already saw that under different choices of assumptions and training setups, RLHF may correspond to different DR-MDP objectives. This is also the case for other common alignment methods, such as the following (which we selected as particularly well known alignment schemes):

- **Reward modeling** (Leike et al., 2018): as we interpret all the techniques we consider in the lens of reward modeling, and such techniques span almost all the objectives in Table 2, it comes as no surprise that depending on the details of reward modeling, it may correspond to different DR-MDP objectives.
- **Cooperative Inverse Reinforcement Learning** (Hadfield-Menell et al., 2016) and **Assistance Games** (Shah et al., 2020): both of these alignment frameworks do not explicitly define the character of the reward learning step, which falls out of the optimization itself. In fact, under certain conditions the CIRL and Assistance Game formalisms reduce to a IRL reward learning step (Hadfield-Menell et al., 2016), but not in others.
- **RLHF** (Christiano et al., 2017; Ouyang et al., 2022): we argued in the previous subsections how under different conditions, RLHF can correspond to the initial reward, myopic reward, final reward, or real-time reward objectives.

There are also many other alignment techniques which we have not considered (Ji et al., 2024), and whose placement among DR-MDP objectives we leave to further work.

G Additional Related Work from Philosophy, Economics, and AI

G.1 Philosophy

Philosophy on welfare under changing preferences. 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, which has some similarities to Griffin (1986)'s criticisms of the "totting-up model" of welfare. Bratman (1987) investigates the role of intention and planning as a coordination mechanism across time for individual decision-making under changing preferences. Bykvist (2006) states that, for judging intertemporal 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 still assumes that these future selves are trustworthy and we value their point of view. Similarly to Bykvist (2006), Paul (2014) and Callard (2018) argue that there is no rational basis for making decisions that change the self. Instead, Pettigrew (2019) builds off of these works and expands on them, challenging the viewpoint that no rational basis is possible for deliberating intertemporal choices, building a theory of individual decision making under changing selves. Paul (2022) remains unconvinced.

Philosophical work on the ethics of influence. While there is a lot of philosophical work on the ethics of influence and manipulation (Noggle, 2020), one line of philosophical work which we have found to be particularly related to ours regards the ethics of "nudging" – a concept that arose in the context of behavioral economics, and refers to institutions trying to influence the behavior and decision-making of groups of individuals outcomes (Thaler & Sunstein, 2008). In our context, the agent performing the nudge can be thought of as the AI, and the person is the one that is nudged. While nudging was originally promoted as a tool to encourage pro-social outcomes, the ethics of its applications have often been contested (Thaler, 2018). There are some philosophical works which take the point of view of an external decision-maker which is assessing whether to perform nudging (Paul & Sunstein, 2019; Pettigrew, 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.3 can be thought of as a generalization of the heuristic proposed by (Pettigrew, 2022), applied to the context of AI.

Work at the intersection of AI and philosophy. The concern with the legitimacy of changes in internal state of humans has also been discussed specifically in connection to AI: in particular, at the

intersection of the two fields there are works on value changes (Ammann, 2024), preference change (Kolodny, 2022; Zhi-Xuan et al., 2024), influence (Bezou-Vrakatseli et al., 2023), and manipulation (Carroll et al., 2023). The concept of "informed preferences" (Gabriel, 2020) 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.2).

G.2 Economics

Shying away from modeling changing preferences. Generally, extending back to at least the 1930s, economists have shied away from analyzing *changing* preferences, for a variety of reasons (George, 2001; Grüne-Yanoff & Hansson, 2009): firstly, preference creation and change have commonly been considered topics that lay outside the scope of economics; second was the conviction of many micro-economists that human preferences ultimately do not change (Stigler & Becker, 1977); a third reason for neglect is the conviction of many macroeconomists that institutional change (relative to changes in individual's preferences), is by far the more important explanatory factor of economic growth. In their seminal paper, Stigler & Becker (1977) 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.

Explaining away changing preferences in terms of hyperbolic discounting. The most established approach to explain any temporal inconsistency of humans is to assume that humans plan using hyperbolic discounting (Loewenstein et al., 2003). Even though hyperbolic discounting may be a good model of people's decision-making in some settings (Benzion et al., 1989; Chabris et al., 2008), this is not the case more broadly (Loewenstein et al., 2003).

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). 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 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—anticipating 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 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").

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."

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).

Algorithmic Amplification in social media. The study of algorithmic amplification in social media (Thorburn et al., 2022; Ribeiro et al., 2023; Huszár et al., 2021; Milli et al., 2023) 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 (Appendix C.1): it's been debated whether one should use random recommendations, a reverse chronological algorithm, competitors' recommenders, or not using any platform at all (Milli et al., 2023).

Performative Prediction and Performative Power. 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. 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. While we recognize that this firm-user misalignment is an additional 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 settings shares various similarities with preference aggregations across multiple individuals, as mentioned in Section 1. This is studied by social choice theory (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. 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).

Representational alignment. Our work can be interpreted in the broader lens of representational alignment (Sucholutsky et al., 2023; Bobu et al., 2024): our DR-MDP formalism has AI systems explicitly represent the fact that our reward evaluations are dependent on our cognitive state, that changes over time and can be influenced by the AI system itself. The hope is that this enables us as designers to better analyze and direct system behavior in such settings.

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 ring similar to the points made by Dobbe et al. (2021).

H Additional Limitations and Discussion

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 (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.2). Moreover, our framework cannot directly express meta-preferences, i.e. the reward functions cannot directly *evaluate transitions between different reward functions*, which may be useful to capture notions of legitimacy of influence and personal autonomy. We leave this to future work.

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.

Toy Problems. Similarly, because the main focus of our work is to clearly describe the problems that one faces with influenceable reward functions, and the challenges with specifying a correct objective, we prioritized the simplicity of our examples for ease of interpretation. An additional challenge that would arise with more complex problems is that of instantiating them realistically without learning reward functions from real users would require developing novel reward learning techniques (as discussed in Appendix A.1).

Misaligned economic pressures. We show that even if AI systems were solely designed with users' welfare in mind, being able to avoid undesirable influence while retaining system capabilities is an unsolved problem. Real-world AI systems will instead be developed under strong economic incentives which are at odds with users' well-being (Susser et al., 2018), giving additional reason to worry about influence from AI systems.

Connection to individual and societal decision-making. Our framework can easily be reinterpreted in the lens of individual (Pettigrew, 2019) and societal (Parkes & Procaccia, 2013) decision making under changing rewards, shedding light on the influence incentives and normative questions implicit in such domains. May also share impossibility results.