Rational Multi-Objective Agents Must Admit Non-Markov Reward Representations

Silviu Pitis[1,3] **Duncan Bailey**[2] **Jimmy Ba**[1,3]

Abstract

This paper considers intuitively appealing axioms for rational, multi-objective agents and derives an impossibility from which one concludes that such agents must admit non-Markov reward representations. The axioms include the Von-Neumann Morgenstern axioms, Pareto indifference, and dynamic consistency. We tie this result to irrational procrastination behaviors observed in humans, and show how the impossibility can be resolved by adopting a non-Markov aggregation scheme. Our work highlights the importance of non-Markov rewards for reinforcement learning and outlines directions for future work.

1 Where Do Rewards Come From?

Sutton's *reward hypothesis* states that the "purpose or goal" of any [rational] agent "can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward)" [39, 35]. Within reinforcement learning (RL), and even machine learning more broadly, this assumption has led to modeling human preferences using a single scalar reward model [8, 38]. A basic result due to Ng and Russell [24] provides some support for this approach: it states that given *any* behavior, there exists a reward function under which that behavior is optimal. Unfortunately, this result speaks only to the *optimal* policy, which we cannot, in general, expect our agents to achieve. Pitis [27] shows value functions computed using such a reward function may misrepresent rational preferences with respect to suboptimal policies, and that this can be resolved by adopting White [44]'s more general task specification with a state-action dependent discount factor. While this more expressive formulation might be understood as supporting a generalized reward hypothesis, important questions remain. Most fundamentally: **where do the rewards come from?**

To make progress on this question, we consider whether multiple simple objectives, represented by relatively simple reward functions (e.g., sparse reward functions or potential functions), can be aggregated into a single, more complex reward function. Indeed, this approach to reward function engineering is commonly used in many (if not most) dense-reward environments; e.g., in OpenAI Gym [5]. In the multi-objective RL literature, it is known as *scalarization* [32]. More generally, this is the object of cardinal social choice, which seeks to aggregate a set of preferences into a single social welfare functional [34]. We note that the problem of aggregating preferences that come from multiple principals or objectives is precisely analogous to the social choice problem generally, so that cardinal utilities are required to resolve Arrow's Impossibility Theorem [3] (e.g., in context of RL, direct aggregation of human preferences or agent policies is theoretically insufficient) (cf. [23]).

We take an axiomatic approach, where both the simple rewards and complex rewards are derived from rational preferences and aggregation is assumed to be Pareto optimal. Interestingly, our chosen axioms lead to an impossibility. We show that if (1) simple rewards are associated with different time preferences (discount functions), and (2) preferences conflict in at least one case, *dynamically consistent, Pareto indifferent preference aggregation is impossible*. This makes reward aggregation a hard problem, even for very simple MDPs: to precisely aggregate rewards with different time

[1]University of Toronto, [2]UC San Diego, [3]Vector Institute. Correspondence to spitis@cs.toronto.edu.

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

preferences we need to adopt a non-Markov reward function or, equivalently, expand the state space. As we aggregate more and more rewards, the complexity of this expansion may be unbounded.

2 The Procrastinator's Peril

To motivate and ground our work, we begin with a numerical example of how the naive aggregation of otherwise rational preferences leads to undesirable behavior. Our example involves repeated procrastination, a phenomenon to which the reader might relate. An agent aggregates two competing objectives: `work` and `play`. At each time step the agent can choose to either `work` or `play`. The pleasure of `play` is mostly from today, and the agent doesn't value future `play` nearly as much as present `play`. On the other hand, the consequences of work are delayed, so that `work` tomorrow is valued approximately as much as `work` today.

Let us model the agent's preferences for `work` and `play` as two separate MDPs, each with state space $\mathcal{S} = \{\mathtt{w}, \mathtt{p}\}$. In the `play` MDP, we have rewards $r(\mathtt{p}) = 0.5$, $r(\mathtt{w}) = 0$ and a discount factor of $\gamma_{\mathtt{play}} = 0.5$. In the `work` MDP, we have rewards $r(\mathtt{p}) = 0$, $r(\mathtt{w}) = 0.3$ and a discount factor of $\gamma_{\mathtt{work}} = 0.9$. One way to combine the preferences for `work` and `play` is to value each trajectory under both MDPs and then add up the values. Not only does this method of aggregation seem reasonable, but it's actually *implied* by some mild and appealing assumptions about preferences (Axioms 1 and 3 in the sequel). Using this approach, the agent assigns values to trajectories as follows:

τ_1	$\mathtt{p}, \mathtt{p}, \mathtt{p}, \mathtt{p}\ldots$	$V(\tau_1) = \sum_t (0.5)^t \cdot 0.5$	$= 1.00$
τ_2	$\mathtt{w}, \mathtt{w}, \mathtt{w}, \mathtt{w}\ldots$	$V(\tau_2) = \sum_t (0.9)^t \cdot 0.3$	$= 3.00$
τ_3	$\mathtt{p}, \mathtt{w}, \mathtt{w}, \mathtt{w}\ldots$	$V(\tau_3) = 0.5 + 0.9 \cdot V(\tau_2)$	$= 3.20$
τ_4	$\mathtt{p}, \mathtt{p}, \mathtt{w}, \mathtt{w}\ldots$	$V(\tau_3) = 0.75 + 0.9^2 \cdot V(\tau_2)$	$= 3.18$

We see that the agent most prefers τ_3: one period (and one period only!) of procrastination is optimal. Thus, the agent procrastinates and chooses to `play` today, planning to `work` from tomorrow onward. Come tomorrow, however, the agent is faced with the same choice decision, and once again puts off `work` in favor of `play`. The process repeats and the agent ends up pursuing the least preferred alternative τ_1. This plainly irrational behavior illustrates our impossibility theorem.

3 Impossibility of Dynamically Consistent, Pareto Indifferent Aggregation

We assume familiarity with Markov Decision Processes (MDPs) [29] and reinforcement learning (RL) [39]. We follow the notation of Pitis [27], denoting an MDP by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$, where $\gamma : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ is a state-action dependent discount factor. We use lowercase letters for generic instances, e.g. $s \in \mathcal{S}$, and denote distributions using a tilde, e.g. $\tilde{s}$. In contrast to standard notation we write both state- and state-action value functions using a unified notation that emphasizes the dependence of each on the future policy: we write $V(s, \pi)$ and $V(s, a\pi)$ instead of $V^\pi(s)$ and $Q^\pi(s, a)$. We extend V to operate on probability distributions of states, $V(\tilde{s}, \Pi) = \mathbb{E}_{s \sim \tilde{s}} V(s, \Pi)$, and we allow for non-stationary, history dependent policies (denoted by uppercase Π, Ω). With this notation, we can understand V as an expected utility function defined over prospects of the form $(\tilde{s}, \Pi)$. For simplicity we assume finite state and action spaces.

3.1 Representing rational preferences

Our work is concerned with the representation of aggregated preferences, where both the aggregation and its individual components satisfy certain axioms of rationality. Our axioms of rationality build on Pitis [27]. We define the objects of preference to be the stochastic processes ("prospects") generated by following (potentially non-stationary and stochastic) policy Π from an state s. Distributions or "lotteries" over these prospects may be represented by (not necessarily unique) tuples of state lottery and policy $(\tilde{s}, \Pi) \in \mathcal{L}(\mathcal{S}) \times \mathbf{\Pi} =: \mathcal{L}(\mathcal{P})$. We write $(\tilde{s}_1, \Pi) \succ (\tilde{s}_2, \Omega)$ if $(\tilde{s}_1, \Pi)$ is strictly preferred to $(\tilde{s}_2, \Omega)$ under preference relation $\succ$. To be rational, we require $\succ$ to satisfy the "VNM axioms" [41], which we capture in Axiom 1:

Axiom 1 (VNM)**.** *For all $\tilde{p}, \tilde{q}, \tilde{r} \in \mathcal{L}(\mathcal{P})$ we have:*

Asymmetry: If $\tilde{p} \succ \tilde{q}$, then not $\tilde{q} \succ \tilde{p}$;
Negative Transitivity: If not $\tilde{p} \succ \tilde{q}$ and not $\tilde{q} \succ \tilde{r}$, not $\tilde{p} \succ \tilde{r}$;
Independence: If $\tilde{p} \succ \tilde{q}$, then $\alpha\tilde{p} + (1-\alpha)\tilde{r} \succ \alpha\tilde{q} + (1-\alpha)\tilde{r}, \forall \alpha \in (0,1]$;
Continuity: If $\tilde{p} \succ \tilde{q} \succ \tilde{r}$, then $\exists\, \alpha, \beta \in (0,1)$ such that $\alpha\tilde{p} + (1-\alpha)\tilde{r} \succ \tilde{q} \succ \beta\tilde{p} + (1-\beta)\tilde{r}$;

where $\alpha\tilde{p} + (1-\alpha)\tilde{q}$ denotes the mixture lottery with $\alpha\%$ chance of $\tilde{p}$ and $(1-\alpha)\%$ chance of $\tilde{q}$.

We require $\succ$ to be *dynamically consistent* [37, 18, 22]:

Axiom 2 (Dynamic consistency)**.** *$(s, a\Pi) \succ (s, a\Omega)$ if and only if $(T(s,a), \Pi) \succ (T(s,a), \Omega)$ where $T(s,a)$ is the distribution over next states after taking action a in state s.*

The axioms produce two key results that we rely on (see Kreps [19] and Pitis [27] for proofs):

Theorem 1 (Expected utility representation)**.** *The relation $\succ$ defined on the set $\mathcal{L}(\mathcal{P})$ satisfies Axiom 1 if and only if there exists a function $V : \mathcal{P} \to \mathbb{R}$ such that, $\forall\, \tilde{p}, \tilde{q} \in \mathcal{L}(\mathcal{P})$:*

$$\tilde{p} \succ \tilde{q} \iff \sum_{z \in supp(\tilde{p})} \tilde{p}(z)V(z) > \sum_{z \in supp(\tilde{q})} \tilde{q}(z)V(z).$$

Another function $V^\dagger$ gives this representation iff $V^\dagger$ is a positive affine transformation of V.

Theorem 2 (Bellman representation)**.** *If $\succ$ satisfies Axioms 1-2*[1] *and V is an expected utility representation of $\succ$, there exist $r : S \times A \to \mathbb{R}$, $\gamma : S \times A \to \mathbb{R}^+$ such that $\forall\, s, a, \Pi$,*

$$V(s, a\Pi) = r(s,a) + \gamma(s,a)V(T(s,a), \Pi).$$

3.2 Representing rational aggregation

We now consider the aggregation of several preferences. These may be the preferences of an agent's several principals or an individual's competing interests. An intuitively appealing axiom for aggregation is Pareto indifference, which says that if each individual preference is indifferent between two alternatives, then so too is the aggregate preference.

Axiom 3 (Pareto indifference)**.** *If $\tilde{p} \approx_i \tilde{q}$ $(\forall i \in \mathcal{I})$, $\tilde{p} \approx_\Sigma \tilde{q}$.*

Here, $\tilde{p} \approx \tilde{q}$ means not $\tilde{p} \succ \tilde{q}$ and not $\tilde{q} \succ \tilde{p}$, $\mathcal{I}$ is a finite index set, and Σ indicates the aggregate relation. There exist stronger variants of Pareto property that require monotonic aggregation [43]. We opt for Pareto indifference to accommodate potentially deviant individual preferences.

If we require individual and aggregate preferences to satisfy Axiom 1 and, jointly, Axiom 3, we obtain a third key result due to [13]. (See Hammond [12] for proof).

Theorem 3 (Harsanyi's representation)**.** *Consider individual preference relations $\{\succ_i; i \in \mathcal{I}\}$ and aggregated preference relation $\succ_\Sigma$, each defined on the set $\mathcal{L}(\mathcal{P})$, that individually satisfy Axiom 1 and jointly satisfy Axiom 3. If $\{V_i; i \in \mathcal{I}\}$ and V_Σ are expected utility representations of $\{\succ_i; i \in \mathcal{I}\}$ and $\succ_\Sigma$, respectively, then there exist real-valued constant c and weights $\{w_i; i \in \mathcal{I}\}$ such that:*

$$V_\Sigma(\tilde{p}) = c + \sum_{i \in \mathcal{I}} w_i V_i(\tilde{p}).$$

3.3 Impossibility result

None of Theorems 1-3 assume all Axioms 1-3. Doing so leads to our main contribution, as follows.

Theorem 4 (Impossibility)**.** *Assume there exist two different policies, Π, Ω, and consider the aggregation of arbitrary individual preference relations $\{\succ_i; i \in \mathcal{I}\}$ defined on $\mathcal{L}(\mathcal{P})$ that individually satisfy Axioms 1-2. There does not exist aggregated preference relation $\succ_\Sigma$ satisfying Axioms 1-3.*

[1] Pitis [27] defined "prospects" as tuples $(\tilde{s}, \Pi)$ and required an additional "Irrelevance of unrealizable actions" axiom. This property is implicit in our redefinition of "prospect" as a stochastic process.

Proof. Fix s, a. Using Theorem 2, choose $\{r_i, \gamma_i\}, r_\Sigma, \gamma_\Sigma$ to represent individual and aggregate preferences over $\mathcal{L}(\mathcal{P})$. W.l.o.g. assume $|\mathcal{I}| = 2$ and shift V_Σ so that by Theorem 3 $\exists \{w_i\}$ for which,

$$\begin{aligned} V_\Sigma(s, a\Pi) - V_\Sigma(s, a\Omega) &= \sum_{i \in \mathcal{I}} w_i V_i(s, a\Pi) - \sum_{i \in \mathcal{I}} w_i V_i(s, a\Omega) \\ &= w_1 \gamma_1(s, a) \left[V_1(T(s, a), \Pi) - V_1(T(s, a), \Omega)\right] + \\ &\qquad w_2 \gamma_2(s, a) \left[V_2(T(s, a), \Pi) - V_2(T(s, a), \Omega)\right]. \end{aligned} \tag{1}$$

Furthermore, plugging in the result of Theorem 3 into Theorem 2 yields,

$$\begin{aligned} V_\Sigma(s, a\Pi) - V_\Sigma(s, a\Omega) &= \gamma_\Sigma(s, a) V_\Sigma(T(s, a), \Pi) - \gamma_\Sigma(s, a) V_\Sigma(T(s, a), \Omega) \\ &= w_1 \gamma_\Sigma(s, a) \left[V_1(T(s, a), \Pi) - V_1(T(s, a), \Omega)\right] + \\ &\qquad w_2 \gamma_\Sigma(s, a) \left[V_2(T(s, a), \Pi) - V_2(T(s, a), \Omega)\right]. \end{aligned} \tag{2}$$

Compare the final RHS of equations (5) and (6) and note that individual preferences $\{\succ_i; i \in \mathcal{I}\}$ may be chosen arbitrarily, so that each term in the square brackets, $V_i(T(s, a), \Pi) - V_i(T(s, a), \Omega)$ is strictly greater than zero. Moreover, we may replace Ω with an α-mixture of Π and Ω (that is, $\alpha\Pi + (1 - \alpha)\Omega$) in the above derivations to obtain equivalent equations where the terms inside the square brackets may assume any value in the closed interval $[0, V_i(T(s, a), \Pi) - V_i(T(s, a), \Omega)]$ and is differentiable with respect to α.

Then, differentiating with respect to α, we obtain the equalities:

$$w_1 \gamma_1(s, a) = w_1 \gamma_\Sigma(s, a) \quad \text{and} \quad w_2 \gamma_2(s, a) = w_2 \gamma_\Sigma(s, a), \tag{3}$$

from which we conclude that $\gamma_\Sigma(s, a) = \gamma_1(s, a) = \gamma_2(s, a)$. But this contradicts our assumption that individual preferences $\{\succ_i; i \in \mathcal{I}\}$ may be chosen arbitrarily, completing the proof. □

An important assumption of Theorem 4 is that the individual preferences $\{\succ_i; i \in \mathcal{I}\}$ may be chosen arbitrarily. This is a natural requirement for aggregation, which is assumed by both Arrow's and Harsanyi's theorems and is commonly referred to as "unrestricted domain". A consequence, critical to the impossibility result and salient in the procrastination example of Section 2, is that the individual preferences may have diverse time preferences (i.e., different discount functions). NB, Theorem 4 generalizes the cases (and so applies) where Π, Ω are stationary, and where γ_1, γ_2, γ_Σ are constants.

The closest results from the economics literature consider consumption streams ($S = \mathbb{R}$) [45, 7, 42]. Within reinforcement learning, equal time preference has been assumed, without justification, when merging MDPs [36, 20].

4 Escaping Impossibility with Non-Markov Rewards

An immediate consequence of Theorem 4 is that any scalarized approach to multi-objective RL [40] is generally insufficient to represent composed preferences. But the implications run deeper: insofar as general tasks compose several objectives, Theorem 4 pushes Sutton's reward hypothesis to its limits. To escape impossibility, the Procrastinator's Peril is suggestive: to be productive, repeat play should not be rewarded. And for this to happen, we must keep track of past play, which suggests that **reward must be non-Markov, even in presence of Markov dynamics**.

The way in which non-Markov rewards (or equivalently, non-Markov utilities/values) escape Theorem 4 is actually quite subtle. Nowhere in the proof of the Theorem, nor in the proofs of Theorems 1, 2, or 3 is the Markov assumption explicitly used. Nor does it obviously appear in any of the Axioms. The Markov assumption *is*, however, invoked in two places. First, to establish comparability between the basic objects of preference—prospects (s, Π)—and second, to extend that comparison set to include "prospects" of the form $(T(s, a), \Pi)$. To achieve initial comparability, Pitis [27] applied a "Markov preference, MP" assumption together with an "original position" construction as follows:

> . . . it is admittedly difficult to express empirical preference over prospects . . . an agent only ever chooses between prospects originating in the same state . . . [Nevertheless,] we imagine a hypothetical state from which an agent chooses between [lotteries] of prospects, denoted by $\mathcal{L}(\mathcal{P})$. We might think of this choice being made from behind a "veil of ignorance" [30]

In other words, to allow for comparisons between prospect (s_1, Π) and (s_2, Π), we prepend some pseudo-state, s_0, and compare prospects $(s_0, s_1\Pi)$ and $(s_0, s_2\Pi)$. Markov preference then lets us cut off the prepended s_0, so that our preferences between (s_1, Π) and (s_2, Π) are cardinal.

Without this construction, however, there is no reason to require relative differences between $V(s_1, *)$ and $V(s_2, *)$ to be meaningful, or to even think about lotteries/mixtures of the two prospects (as done in Axiom 1). Letting go of this ability to compare prospects starting in different states means that Theorem 1 is applicable only to sets of prospects with matching initial states.

Though this does not directly affect the conclusion of Theorem 2, $T(s, a)$ in $V(T(s, a), \Pi)$ includes a piece of history, (s, a), and can no longer be computed as $\mathbb{E}_{s' \sim T(s,a)} V(s', \Pi)$. Instead, since the agent is not choosing between prospects of form (s', Π) but rather (abusing notation) prospects of form (sas', Π), the expectation should be computed as $\mathbb{E}_{s' \sim T(s,a)} V(sas', \Pi)$.

The inability to compare prospects starting in different states also affects the conclusion of Theorem 3, which implicitly uses such comparisons to find constant coefficients w_i that apply everywhere in the original $\mathcal{L}(\mathcal{P})$. Relaxing Harsanyi's theorem to not make inter-state comparisons results in coefficients that are state dependent when aggregating utilities of prospects in the original $\mathcal{L}(\mathcal{P})$, and history dependent when aggregating the historical prospects—e.g., of the form (sas', Π)—as is now required to compute the non-Markov value $V(T(s, a), \Pi)$ that appears in Theorem 2.

Allowing the use of history dependent coefficients in the aggregation of $V_i(T(s, a), \Pi)$ resolves the impossibility. The following result shows that given state dependent coefficients $w_i(s)$, we can always construct history dependent coefficients $w_i(sas')$ that allow for dynamically consistent aggregation that satisfies all axioms. In the statement of the theorem, we use $\mathcal{L}(\mathcal{P}_y)$ to denote the set of prospect lotteries starting with history y.

Theorem 5 (Possibility)**.** *Consider the aggregation of arbitrary individual preference relations* $\{\succ_i^y; i \in \mathcal{I}\}$ *defined on their respective* $\mathcal{L}(\mathcal{P}_y)$*, for each history* y*, that individually satisfy Axioms 1-2. There exist aggregated preference relations* $\{\succ_\Sigma^y\}$ *each satisfying Axioms 1-3.*

In particular, ***given*** $s, a, V_\Sigma, \{V_i\}, \{w_i(s)\}$*,* ***where (A)*** V_Σ *and each* V_i *individually satisfy Axioms 1-2 on* $\mathcal{L}(\mathcal{P}_y), \forall y$*,* ***and (B)*** $V_\Sigma(s, \Pi) = \sum_i w_i(s) V_i(s, a\Pi))$*,* ***then****, choosing*

$$w_i(sas') = w_i(sa) = w_i(s)\gamma_i(s, a) \quad \textit{for all } i, s' \tag{4}$$

implies that $V_\Sigma(T(s, a), \Pi) \propto \sum_i w_i(sa) V_i(T(s, a), \Pi)$ *so that the aggregated preferences* $\{\succ_\Sigma^{sas'}\}$ *satisfy Axiom 3. Unrolling this result produces a set of constructive, history dependent weights* $\{w_i(y)\}$ *that satisfy Axiom 3 for all histories* $\{y\}$.

Proof. We follow the proof of Theorem 4. Fix s, a. Using Theorem 2, choose $\{r_i, \gamma_i\}, r_\Sigma, \gamma_\Sigma$ to represent individual and aggregate preferences over $\mathcal{L}(\mathcal{P}_s)$ and $\mathcal{L}(\mathcal{P}_{sas'})$. W.l.o.g. assume $|\mathcal{I}| = 2$ and shift V_Σ so that by Theorem 3 $\exists \{w_i(s)\}$ for which,

$$\begin{aligned}
V_\Sigma(s, a\Pi) - V_\Sigma(s, a\Omega) &= \sum_{i \in \mathcal{I}} w_i(s) V_i(s, a\Pi) - \sum_{i \in \mathcal{I}} w_i(s) V_i(s, a\Omega) \\
&= w_1(s)\gamma_1(s, a) \left[V_1(T(s, a), \Pi) - V_1(T(s, a), \Omega)\right] + \\
&\qquad w_2(s)\gamma_2(s, a) \left[V_2(T(s, a), \Pi) - V_2(T(s, a), \Omega)\right].
\end{aligned} \tag{5}$$

Furthermore, plugging in the result of Theorem 3 into Theorem 2 yields,

$$\begin{aligned}
V_\Sigma(s, a\Pi) - V_\Sigma(s, a\Omega) &= \gamma_\Sigma(s, a) V_\Sigma(T(s, a), \Pi) - \gamma_\Sigma(s, a) V_\Sigma(T(s, a), \Omega) \\
&= \gamma_\Sigma(s, a) \mathbb{E}_{s'} \left[V_\Sigma(sas', \Pi)\right] - \gamma_\Sigma(s, a) E_{s'} \left[V_\Sigma(sas', \Omega)\right] \\
&= \gamma_\Sigma(s, a) E_{s'} \left[w_1(sa) V_1(T(s, a), \Pi) - w_1(sa) V_1(T(s, a), \Omega)\right] + \\
&\qquad \gamma_\Sigma(s, a) E_{s'} \left[w_2(sa) V_2(T(s, a), \Pi) - w_2(sa) V_2(T(s, a), \Omega)\right]. \\
&= w_1(sa)\gamma_\Sigma(s, a) \left[V_1(T(s, a), \Pi) - V_1(T(s, a), \Omega)\right] + \\
&\qquad w_2(sa)\gamma_\Sigma(s, a) \left[V_2(T(s, a), \Pi) - V_2(T(s, a), \Omega)\right].
\end{aligned} \tag{6}$$

As in Theorem 4, we define and differentiate with respect to an α-mixture of Π and Ω to obtain:

$$w_1(s)\gamma_1(s, a) = w_1(sa)\gamma_\Sigma(s, a) \quad \text{and} \quad w_2(s)\gamma_2(s, a) = w_2(sa)\gamma_\Sigma(s, a), \tag{7}$$

from which we conclude that:

$$\frac{w_2(sa)}{w_1(sa)} = \frac{w_2(s)\gamma_2(s, a)}{w_1(s)\gamma_1(s, a)} \tag{8}$$

This shows the existence of weights $w_i(sa)$, unique up to a constant scaling factor, for which $V_\Sigma(T(s,a),\Pi) \propto \sum_i w_i(sa)V_i(T(s,a),\Pi)$, that apply regardless of how individual preferences are chosen or aggregated at s. Unrolling the result completes the proof. □

From Theorem 5 we obtain a rather elegant result: rational aggregation over time discounts the aggregation weights assigned to each individual value function its respective discount factor. In the Procrastinator's Peril, for instance, where we started with $w_{\texttt{p}}(s) = w_{\texttt{w}}(s) = 1$, at the initial (and only) state s, we have $w_{\texttt{p}}(s\texttt{p}s) = 0.5$ and $w_{\texttt{w}}(s\texttt{p}s) = 0.9$. With these non-Markov aggregation weights, you can verify that (1) the irrational procrastination behavior is solved, and (2) the aggregated rewards for `work` and `play` are now non-Markov.

5 Non-Markov Rewards in Reinforcement Learning

The necessity of non-Markov rewards was demonstrated in other settings by Abel et al. [1], and would also arise simply from Axioms 1-2 if state-action dependent discounts could not be considered [27]. Though several papers explicitly consider RL with non-Markov rewards [10, 31], this is usually motivated by task compression rather than necessity, and the majority of the RL literature restricts itself to Markov rewards models. We note that many popular exploration strategies implicitly use non-Markov rewards [17, 25], and that certain work has also considered non-Markov discount factors [9, 33].

Conveying Non-Markov Reward The basic approach to dealing with non-Markov reward is to expand the state space in such a way that reward becomes Markov [10]. Several ways of doing so have been proposed, including Reward Machines [6] and the Split-MDP [2]. However, the expanded reward state space is exponential in the number of acceptable policies and may be unbounded in the amount of simple reward functions that are aggregated [2].

Enriched State Spaces Several specialized subtopics in RL use similarly enriched state spaces, which can be related to non-Markov rewards. Perhaps the best known is the use of MDPs to address partially observed problems, by solving an MDP over "belief states" [21], often represented as a function of a compressed version of the history [16]. A similar preference aggregation problem may arise in multi-agent RL, and admits a similar treatment to POMDPs whilst also involving a more explicit aggregation of preferences from the individual agents [11]. In multi-goal RL [28] as well as works on fast multi-task transfer [4], the reward representation is held separate from the state dynamics model, which allows for strategies such as ex-post reward relabeling [15] and ex-ante intrinsic reward setting [26]. These strategies might be adapted to improve the learning speed of aggregation-based agents that use non-Markov rewards.

6 Conclusion

The main contribution of this paper is an impossibility result from which one concludes that non-Markov rewards are likely necessary for reinforcement learning agents that either (1) pursue multiple simultaneous objectives, and/or (2) serve multiple principals with differing preferences. We suspect this will be the case for any advanced RL agent. To accurately align such agents with diverse human preferences and minimize x-risks (Appendix A), we need to endow them with the capacity to represent non-Markov reward (Section 5). Future work should quantify the inefficiency of using Markov representations where non-Markov rewards are necessary, explore alternatives such as commitment-based strategies that leave non-Markov rewards implicit, and consider the explicit construction of aggregated non-Markov reward functions.

References

[1] Abel, D.; Dabney, W.; Harutyunyan, A.; Ho, M. K.; Littman, M.; Precup, D.; and Singh, S. 2021. On the expressivity of markov reward. *Advances in Neural Information Processing Systems* 34:7799–7812.

[2] Abel, D.; Barreto, A.; Bowling, M.; Dabney, W.; Hansen, S.; Harutyunyan, A.; Ho, M.; Kumar, R.; Littman, M.; Precup, D.; and Singh, S. 2022. Expressing non-markov reward to a markov agent. In *RLDM*, 1–5.

[3] Arrow, K. J.; Sen, A.; and Suzumura, K. 2010. *Handbook of social choice and welfare*, volume 2. Elsevier.

[4] Barreto, A.; Dabney, W.; Munos, R.; Hunt, J. J.; Schaul, T.; van Hasselt, H. P.; and Silver, D. 2017. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems* 30.

[5] Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

[6] Camacho, A.; Icarte, R. T.; Klassen, T. Q.; Valenzano, R. A.; and McIlraith, S. A. 2019. Ltl and beyond: Formal languages for reward function specification in reinforcement learning. In *IJCAI*, volume 19, 6065–6073.

[7] Chambers, C. P., and Echenique, F. 2018. On multiple discount rates. *Econometrica* 86(4):1325–1346.

[8] Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*.

[9] Fedus, W.; Gelada, C.; Bengio, Y.; Bellemare, M. G.; and Larochelle, H. 2019. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*.

[10] Gaon, M., and Brafman, R. 2020. Reinforcement learning with non-markovian rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 3980–3987.

[11] Gmytrasiewicz, P. J., and Doshi, P. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research* 24:49–79.

[12] Hammond, P. J. 1992. Harsanyi's utilitarian theorem: A simpler proof and some ethical connotations. In *Rational Interaction*. Springer. 305–319.

[13] Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy* 63(4):309–321.

[14] Hendrycks, D., and Mazeika, M. 2022. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*.

[15] Kaelbling, L. P. 1993. Learning to achieve goals. In *IJCAI*, volume 2, 1094–8. Citeseer.

[16] Kapturowski, S.; Ostrovski, G.; Quan, J.; Munos, R.; and Dabney, W. 2018. Recurrent experience replay in distributed reinforcement learning. In *International conference on learning representations*.

[17] Kolter, J. Z., and Ng, A. Y. 2009. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, 513–520.

[18] Kreps, D. M., and Porteus, E. L. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica: journal of the Econometric Society* 185–200.

[19] Kreps, D. 1988. *Notes on the Theory of Choice*. Westview press.

[20] Laroche, R.; Fatemi, M.; Romoff, J.; and van Seijen, H. 2017. Multi-advisor reinforcement learning. *arXiv preprint arXiv:1704.00756*.

[21] Littman, M. L. 2009. A tutorial on partially observable markov decision processes. *Journal of Mathematical Psychology* 53(3):119–125.

[22] Machina, M. J. 1989. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature* 27(4):1622–1668.

[23] Muandet, K. 2022. Impossibility of collective intelligence. *arXiv preprint arXiv:2206.02786*.

[24] Ng, A. Y., and Russell, S. J. 2000. Algorithms for inverse reinforcement learning. In *The Seventeenth International Conference on Machine Learning*, 663–670.

[25] Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, 2778–2787. PMLR.

[26] Pitis, S.; Chan, H.; Zhao, S.; Stadie, B.; and Ba, J. 2020. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. In *International Conference on Machine Learning*, 7750–7761. PMLR.

[27] Pitis, S. 2019. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7949–7956.

[28] Plappert, M.; Andrychowicz, M.; Ray, A.; McGrew, B.; Baker, B.; Powell, G.; Schneider, J.; Tobin, J.; Chociej, M.; Welinder, P.; et al. 2018. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*.

[29] Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[30] Rawls, J. 2009. *A theory of justice: Revised edition*. Harvard university press.

[31] Rens, G.; Raskin, J.-F.; Reynoaud, R.; and Marra, G. 2020. Online learning of non-markovian reward models. In *Proceedings of the Thirteenth International Conference on Agents and Artificial Intelligence*. Scitepress.

[32] Roijers, D. M.; Vamplew, P.; Whiteson, S.; and Dazeley, R. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* 48:67–113.

[33] Schultheis, M.; Rothkopf, C. A.; and Koeppl, H. 2022. Reinforcement learning with non-exponential discounting. In *Advances in neural information processing systems*.

[34] Sen, A. 2017. *Collective Choice and Social Welfare: Expanded Edition*. Penguin UK.

[35] Silver, D.; Singh, S.; Precup, D.; and Sutton, R. S. 2021. Reward is enough. *Artificial Intelligence* 299:103535.

[36] Singh, S. P., and Cohn, D. 1998. How to dynamically merge markov decision processes. In *Advances in neural information processing systems*, 1057–1063.

[37] Sobel, M. J. 1975. Ordinal dynamic programming. *Management science* 21(9):967–975.

[38] Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33:3008–3021.

[39] Sutton, R. S., and Barto, A. G. 2018. *Reinforcement Learning: An Introduction (in preparation)*. MIT Press.

[40] Van Moffaert, K.; Drugan, M. M.; and Nowé, A. 2013. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 191–199. IEEE.

[41] Von Neumann, J., and Morgenstern, O. 1953. *Theory of games and economic behavior*. Princeton university press.

[42] Weitzman, M. L. 2001. Gamma discounting. *American Economic Review* 91(1):260–271.

[43] Weymark, J. A. 1993. Harsanyi's social aggregation theorem and the weak pareto principle. *Social choice and welfare* 10(3):209–221.

[44] White, M. 2017. Unifying task specification in reinforcement learning. In *The Thirty-fourth International Conference on Machine Learning*.

[45] Zuber, S. 2011. The aggregation of preferences: can we ignore the past? *Theory and decision* 70(3):367–384.

A X-Risk Sheet

We analyze existential risk using the X-Risk Sheet Template of Hendrycks and Mazeika [14].

Individual question responses do not decisively imply relevance or irrelevance to existential risk reduction. Do not check a box if it is not applicable.

A.1 Long-Term Impact on Advanced AI Systems

In this section, please analyze how this work shapes the process that will lead to advanced AI systems and how it steers the process in a safer direction.

1. **Overview.** How is this work intended to reduce existential risks from advanced AI systems?

   Answer: We showed that non-Markov reward functions can resolve proxy misspecification of AI systems that would otherwise arise viaimproper reward aggregation. Starting from several intuitive axioms of rational aggregation, an impossibility arising from multiple objectives is shown. Thus we conclude that, in order to reduce reward hacking and proxy misspecification, non-Markov rewards are needed and can be defined using a multiple mechanisms. This conclusion could help focus reinforcement learning research on questions that matter, rather than have researchers waste cycles on solutions that apply only to Markov rewards. Our conclusion applies to any multi-objective agent, whose individual objectives may possess multiple time preferences. We suspect that this will include all advanced AI systems, as they are bound to have multiple human principals with varying objectives, and humans have been shown to possess varying time preference.

   Insofar as RL agents relate to existential risk, reward function design is paramount, as reward functions are the fundamental tool for agent alignment. To better align AI preferences (defined by a reward function and discount factor) with human preferences, we must better understand the theoretical insufficiencies of current reward representations, which is what this work aims to do.

2. **Direct Effects.** If this work directly reduces existential risks, what are the main hazards, vulnerabilities, or failure modes that it directly affects?

   Answer: As noted above, our conclusions, if acted upon, could directly reduce proxy misspecification, AI misgeneralization, and reward hacking. Risks of deceptive alignment, emergent behaviors and goals, maliciously steered AI and colluding AIs emerging from a Markov reward might be better understood and mechanisms to reduce this risk are outlined in our paper.

3. **Diffuse Effects.** If this work reduces existential risks indirectly or diffusely, what are the main contributing factors that it affects?

   Answer: This work could improve incentive structures, standards, defense in depth, and reduce the potential for human error via reward alignment.

4. **What's at Stake?** What is a future scenario in which this research direction could prevent the sudden, large-scale loss of life? If not applicable, what is a future scenario in which this research direction be highly beneficial?

   Answer: In the case that humans want an AI to pursue an objective that is multi-faceted (i.e. combines multiple simple objectives, or the joint objective of a multitude of agents), a well defined reward function may not exist, and a naive aggregation may incentivize unintended behavior. This can have catastrophic effects if the AI is given power over human life. Understanding the foundation and limitation of reward functions provides important insight into task specification and human value alignment.

5. **Result Fragility.** Do the findings rest on strong theoretical assumptions; are they not demonstrated using leading-edge tasks or models; or are the findings highly sensitive to hyperparameters? □

6. **Problem Difficulty.** Is it implausible that any practical system could ever markedly outperform humans at this task? □

7. **Human Unreliability.** Does this approach strongly depend on handcrafted features, expert supervision, or human reliability? ☒

8. **Competitive Pressures.** Does work towards this approach strongly trade off against raw intelligence, other general capabilities, or economic utility? ☒

A.2 Safety-Capabilities Balance

In this section, please analyze how this work relates to general capabilities and how it affects the balance between safety and hazards from general capabilities.

9. **Overview.** How does this improve safety more than it improves general capabilities?

Answer: The discussed solutions to reward aggregation increases pre-deployment costs (e.g. constructing a split-MDP is more intensive on the reward designer) and improves safety. It also increases the interpretability of generalization results occurring from a single reward structure. Our work has little impact on general capabilities (except to the extent that accurate aggregation of preferences is considered a capability).

10. **Red Teaming.** What is a way in which this hastens general capabilities or the onset of x-risks?

    Answer: This work does not further capabilities; it mostly comments on the theoretical basis of differing incentive structures for agents.

11. **General Tasks.** Does this work advance progress on tasks that have been previously considered the subject of usual capabilities research? ☒

12. **General Goals.** Does this improve or facilitate research towards general prediction, classification, state estimation, efficiency, scalability, generation, data compression, executing clear instructions, helpfulness, informativeness, reasoning, planning, researching, optimization, (self-)supervised learning, sequential decision making, recursive self-improvement, open-ended goals, models accessing the Internet, or similar capabilities? ☒

13. **Correlation With General Aptitude.** Is the analyzed capability known to be highly predicted by general cognitive ability or educational attainment? ☒

14. **Safety via Capabilities.** Does this advance safety along with, or as a consequence of, advancing other capabilities or the study of AI? ☒

A.3 Elaborations and Other Considerations

15. **Other.** What clarifications or uncertainties about this work and x-risk are worth mentioning?

    Answer: Relating to Q5, all the assumptions/axioms imposed on preferences over prospects are deemed to be rational and normative, therefore not believed to be strong theoretical assumptions.