Settling the Reward Hypothesis

Michael Bowling [*1] **John D. Martin** [*1 2] **David Abel** [3] **Will Dabney** [3]

Abstract

The *reward hypothesis* posits that, "all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)." We aim to fully settle this hypothesis. This will not conclude with a simple affirmation or refutation, but rather specify completely the implicit requirements on goals and purposes under which the hypothesis holds.

1. Introduction

The *reward hypothesis* posited by Sutton states that, "all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)." (Sutton, 2004; Sutton & Barto, 2018; Littman, 2017). This statement takes on considerable import if one also accepts McCarthy's claim that "Intelligence is the computational part of the ability to achieve goals in the world." (McCarthy, 1998). Together these two statements offer a sort of *sufficiency* to the study of reinforcement learning (RL), whose agents learn to achieve goals through the maximization of expected future rewards. They imply that to succeed at building AI, it is sufficient to succeed at solving RL.[1]

Silver et al. (2021) propose the related, *reward-is-enough* hypothesis, which posits that "intelligence, and all of its associated abilities, can be understood as subserving the maximization of reward." While the two hypotheses are of course deeply connected, we emphasize that our focus is on Sutton's earlier *reward hypothesis*.

Sutton's original hypothesis provides an informal starting point from which to question the expressivity of reward. In this vein, Abel et al. (2021) grounded the notion of "goals and purposes" as an ordering over policies and explored whether a Markov reward[2] function could express these orderings. They provided examples showing that a Markov reward is unable to express every such ordering. Their analysis also reveals how using a behavioral definition of goals can sometimes lead to unsatisfying conclusions. In one example ("steady-state type" failures), an agent needs to experience unrealizable outcomes to achieve their goal. When viewed through the lens of McCarthy's definition of intelligence, it seems that a behavioral conception of goals deflates the role that computation performs—to one of simply executing a goal's defining policy, or re-expressing the policy in a different form. Surely intelligence involves more meaningful computation than this?

Shakerinava & Ravanbakhsh (2022) take a different approach. They ground goals and purposes in preference relations over (distributions of) state-trajectories in a controlled Markov process. In the same spirit of von Neumann-Morgenstern (vNM) utility axioms (von Neumann & Morgenstern, 1953), they propose axioms on the preference relation (including the vNM rationality axioms) which are necessary and sufficient for the preference to be expressed with a Markov reward. Indeed, Shakerinava & Ravanbakhsh (2022) build on work by Pitis (2019) that first analyzed standard objectives of RL from the perspective of decision theory. The work of Pitis can be viewed from two complementary perspectives. First, Pitis provides a normative account for why we should embrace a state-action dependent discount factor, as developed by White (2017): A fixed discount cannot capture all preferences we might consider rational. Second, Pitis presents three axioms on top of the vNM axioms that characterize the conditions under which a state-action dependent discount factor can be viewed as rational.

Our work builds off this pair of insightful approaches by starting with preferences over histories. We abandon strictly Markov processes to consider general stochastic environments and policies in line with recent work by Dong et al.

[*]Equal contribution [1]Amii, University of Alberta [2]Intel Labs [3]DeepMind. Correspondence to: John D. Martin <john.martin@intel.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]Furthermore, it is possible these two claims show that RL is necessary as well as sufficient. If some artificial system were to be able to achieve all that we mean by goals and purposes, then such a system would have to—at least implicitly—maximize expected cumulative reward. In other words, there must be a reduction between the RL problem and the problem the system solves.

[2]They take a Markov reward function to be one that only depends on the most recent experience of the agent.

(2021), Lu et al. (2021), and early work on general RL (Lattimore et al., 2013; Lattimore, 2014; Leike, 2016; Majeed, 2021). We introduce a new axiom that generalizes previous axioms from Shakerinava & Ravanbakhsh (2022) and accommodates the discounted reward, average reward (Mahadevan, 1996), and episodic settings. Our approach posits "goals and purposes" as preceding environment dynamics, giving space for the agent's computational role of learning representations and behavior necessary to accomplish a goal. Using our new axiom along with the standard vNM rationality axioms, we provide a treatment of the reward hypothesis in both the setting that goals are the subjective desires of the agent and in the setting where goals are the objective desires of an agent designer. Altogether, our account does not give a simple affirmation or refutation of the reward hypothesis, but rather aims to completely specify the implicit requirements on goals and purposes under which the hypothesis holds.

2. The Reward Hypothesis

As we aim to settle the reward hypothesis, the first step in doing so is to formalize what it claims, and to do so in as much generality as possible. We do this by stating a series of assumptions for each of the phrases in the claim.

2.1. Goals as Preferences

We ground "all of what we mean by goals and purposes" with a binary preference relation expressing preference for one outcome over another.

The core of agent interaction is the cycle of repeatedly observing the environment and taking action to affect the environment. Let $\mathcal{O}$ be a finite set of observations, and $\mathcal{A}$ a finite set of actions.[3] A history is then a sequence $o_1, a_1, o_2, a_2, \ldots$ with ε as the empty history of zero length. We define the set of histories of length $n \in \mathbb{N}_{\geq 0}$ as $\mathcal{H}_n \stackrel{\text{def}}{=} (\mathcal{O} \times \mathcal{A})^n$, and all finite length histories as $\mathcal{H} \stackrel{\text{def}}{=} \bigcup_{n=1}^{\infty} \mathcal{H}_n$. For $h \in \mathcal{H}$ and transition $t \in \mathcal{O} \times \mathcal{A}$, let $t \cdot h \in \mathcal{H}$ be the history with t prepended to h.

In deterministic settings, preferences are over histories. However, when the environment or agent behavior are stochastic, we consider distributions over histories, $\Delta(\mathcal{H})$.[4] Given $A, B \in \Delta(\mathcal{H})$ and $p \in [0, 1]$, let $pA + (1-p)B \in \Delta(\mathcal{H})$ be the distribution that samples a history from A with probability p and B with $(1-p)$. For $A \in \Delta(\mathcal{H})$ and $t \in \mathcal{O} \times \mathcal{A}$, let $t \cdot A \in \Delta(\mathcal{H})$, be the distribution where t is prepended to a history sampled from A.

[3] We assume $\mathcal{O}$ and $\mathcal{A}$ are finite, but suspect the results generalize to the case where they are simply countable sets.

[4] We use $\Delta(\mathcal{S})$ to refer to the set of all probability distributions with finite support over a countable set $\mathcal{S}$.

The reward hypothesis posits a "received scalar signal". This could mean that the posited scalar reward signal is present in the agent's received observation, or can be computed by the agent from it. Alternatively, it could mean that there is an additional scalar signal provided to the agent by an external observer. We call the first setting *subjective goals*, as the posited reward signal can be constructed from the agent's subjective observation, and the latter case correspondingly *objective goals*. We first develop our main result with the subjective goals setting, but will later broaden the result to objective goals.

Assumption 1 (Subjective Goals)**.** *"All of what we mean by goals and purposes" can be expressed as a binary preference relation on distributions over finite histories. For $A, B \in \Delta(\mathcal{H})$, we write $A \succsim B$ if A is weakly preferred to B, meaning that either A is strictly preferred B, or the two are indifferently preferred, which we denote $A \sim B$.*

Notice that our notion of "goals and purposes" make no reference to the environment. Goals are stated as desirable histories, whereas the environment will act to constrain what histories and distributions over histories are possible. An agent's behavior, along with the environment, then induces a distribution over histories. Formally, given an environment $e : \mathcal{H} \to \Delta(\mathcal{O})$ and policy $\pi : \mathcal{H} \times \mathcal{O} \to \Delta(\mathcal{A})$, let D_n^π be the distribution over $\mathcal{H}_n$ induced by e and π.

$$D_n^\pi \stackrel{\text{def}}{=} \Pr\left[o_1, a_1, \ldots, o_n, a_n | e, \pi\right] = \prod_{i=1}^{n} e(o_i | o_1, a_1, \ldots a_{i-1}) \pi(a_i | o_1, a_1, \ldots, o_i).$$

While D_n^π depends on the environment e, we do not parameterize it as such since e typically is fixed throughout. We assume preferences over agent policies are then consistent with the distributions over histories that they induce. When comparing policies, we write $\pi_1 \succsim_g \pi_2$ to mean π_1 is weakly preferred to π_2 under the goal g.

Assumption 2 (Policy Preferences)**.** *We weakly prefer $\pi_1 \succsim_g \pi_2$ in e if and only if there exists N such that $D_n^{\pi_1} \succsim D_n^{\pi_2}$ for all $n \geq N$.*

This notion of eventually preferring one policy's history distribution to another allows us the generality of goals and purposes that can be achieved in a defined time frame as well as those of a continuing nature. This assumption is just one simple way of handling infinite sequences, but it is not the only one. For instance, Sobel (1975) and Pitis (2019) propose alternative resolutions (horizon continuity and countable transitivity—see details in cited resources).

2.2. Maximizing Cumulative Sums

We now consider what the reward hypothesis says about these goals and purposes. First, let us examine what is entailed by the "maximization of the cumulative sum" of a

scalar reward. This is clearly the domain of reinforcement learning, of which this problem takes a number of forms, including episodic total reward with absorbing states, infinite sum of discounted rewards, and average reward. We want to unify all of these formalisms in what could be meant by maximizing cumulative sums of rewards, and do so employing White's generalization of transition-dependent discounting (White, 2017). When comparing policies under a reward r, we write $\pi_1 \succsim_r \pi_2$.

Assumption 3 (Cumulative Sum of Rewards)**.** *The "maximization of the expected value of the cumulative sum of a received scalar signal (reward)" means that there is a reward function $r : \mathcal{O} \times \mathcal{A} \to \mathbb{R}$ and transition-dependent discount function $\gamma : \mathcal{O} \times \mathcal{A} \to [0, 1]$, such that we weakly prefer $\pi_1 \succsim_r \pi_2$ under our reward if and only if there exists N such that $V_n^{\pi_1} \geq V_n^{\pi_2}$ for all $n \geq N$, where*

$$V_n^\pi \stackrel{\text{def}}{=} E\left[\sum_{i=1}^{n}\left(\prod_{j=1}^{i-1}\gamma(O_j, A_j)\right) r(O_i, A_i)\middle|\pi, e\right]. \quad (1)$$

Notice how (1) simultaneously captures objectives for the average reward, discounted reward, and episodic settings. If $\gamma(t) = 1$ for some o, a pair t, the objective corresponds to a typical total reward or average reward setting (even if the sum of the reward is not bounded). If $\gamma(t) = \gamma < 1$ is constant for all t, then the objective corresponds to a discounted reward objective. If $\gamma(t) = 0$ infinitely often then the objective can correspond to an episodic setting. In the Appendix, we expand on the average reward case and show how the notion of the expected cumulative sum eventually being larger allows us to capture multiple kinds of optimality.

We can now state what we take the reward hypothesis to mean under all of these assumptions.

Assumption 4 (Reward Hypothesis)**.** *What the reward hypothesis means by "well thought of" is that for any preference relation on distributions of histories there exists r and γ such that $\pi_1 \succsim_g \pi_2$ under the goal g iff $\pi_1 \succsim_r \pi_2$.*

In what follows, we explore if this is true or false, or more precisely, what might be required of our preference relation for the reward hypothesis to hold.

3. Rationality Axioms

The vNM axioms provide necessary and sufficient conditions for a preference relation to be expressible as the expectation of some scalar-valued function of outcomes. The "expected value of the cumulative sum" of rewards is central to the reward hypothesis, so we present these axioms here along with the corresponding vNM utility theorem. In the statement of these and additional axioms, $\mathcal{H}$ is any set of finite length sequences of some countable set of transitions T (e.g., T may be $\mathcal{O} \times \mathcal{A}$ as in the subjective goals case of Assumption 1).

We next state each of the four vNM axioms alongside some brief intuition.

Axiom 1 (Completeness)**.** *For all $A, B \in \Delta(\mathcal{H})$, $A \succsim B$ or $B \succsim A$ (or both, if $A \sim B$).*

Completeness requires that the preference ordering make *some* judgment about any pair of distributions. Note that the preference *could* simply convey indifference: we might be equally satisfied with an apple and a banana. This is distinct from having no preference at all (Chang (2015) discusses the incomparability of virtues like "justice" and "mercy"). Note that there are alternative sets of axioms that simply remove completeness, as developed by Aumann (1962).

Axiom 2 (Transitivity)**.** *For all $A, B, C \in \Delta(\mathcal{H})$, if $A \succsim B \succsim C$, then $A \succsim C$.*

Transitivity is relatively straightforward: no coherent goal can involve cyclical preferences.

Axiom 3 (Independence)**.** *For all $A, B, C \in \Delta(\mathcal{H})$ and $p \in (0, 1)$, $A \succsim B$ if and only if*

$$pA + (1 - p)C \succsim pB + (1 - p)C$$

Independence was historically viewed as an unlikely axiom, and one that led to skepticism about vNM's initial results (Machina, 1990). However, it does convey a relatively powerful intuition, once it is unpacked. Consider the following example. Suppose we must choose between an Apple, a Banana, or Chocolate. We are asked: (1) Do you prefer one Apple to one Banana? and (2) Consider two coins of the same weight which, when flipped can yield {H: Apple, T: Chocolate}, and {H: Banana, T: Chocolate}—which coin do you prefer to flip? Independence requires that for *all* coin weights, your answer to (1) and (2) must be the same. In other words, if you truly prefer an Apple to a Banana, then the *chance* of procuring the Apple and Banana should not change this preference (when the other alternatives are the same). Independence is deeply connected to many alternative objectives in RL such as risk-aversion and multi-objective RL, which we discuss in more detail shortly.

Axiom 4 (Continuity)**.** *For all $A, B, C \in \Delta(\mathcal{H})$ if $A \succsim B \succsim C$, then there exists $p \in [0, 1]$ such that,*

$$pA + (1 - p)C \sim B$$

Continuity demands the existence of a break-even point when one becomes indifferent to a mixture of outcomes that are individually more or less preferred. At what precise coin

weight does one become indifferent between a Banana and a distribution of Apple and Chocolate?

These four axioms comprise the typical account of rational preferences under vNM's theory. We next state the classical vNM result.

Theorem 3.1 (von Neumann-Morgenstern Utility Theorem)**.** *A binary preference relation on $\Delta(\mathcal{H})$ satisfies axioms 1-4 if and only if there exists a utility function $u : \Delta(\mathcal{H}) \to \mathbb{R}$ such that,*

1. $\forall A, B \in \Delta(\mathcal{H}) \quad A \succsim B \Leftrightarrow u(A) \geq u(B)$,
2. $\forall (\sum_i p_i h_i) \in \Delta(\mathcal{H}) \quad u(\sum_i p_i h_i) = \sum_i p_i u(h_i)$,

and u is unique up to positive affine transformations.

In other words, there exists a utility function whose expectation for any distribution over histories is consistent with the preference relation.

These rationality axioms are sufficient for our interpretation of the reward hypothesis (Assumption 4) to trivially hold with no further assumptions. Simply define the "received" reward for experiencing transition t following history h as the change in utility from appending transition t to h, written as $r(t; h) = u(h \cdot t) - u(h)$. However, this reward function is not, in general, computable from the agent's received observation as it depends on the entire history. This definition of reward function may not even be computable by a bounded agent or a bounded external observer as, in general, it requires complete memory of the history. We will need to add an additional axiom for the reward hypothesis to hold for a Markov reward; that is, a reward that is received or computable from the agent's most immediate experience.

4. A New Axiom: Temporal γ-Indifference

Here we derive a new axiom that ensures the existence of Markov rewards and implicitly specifies the type of objective faced by an agent. We start from an observation about how preferences can still encode statements about magnitude, even without reducing them to a scalar value.

Let $A, B, C, D \in \Delta(\mathcal{H})$ with $A \succsim B$ and $C \succsim D$. Suppose we wanted to state that A is preferred over B *by the same amount* as C is preferred over D. If we could encode preferences as utilities, then we could write

$$\begin{aligned} u(A) - u(B) &= u(C) - u(D), \\ u(A) + u(D) &= u(B) + u(C), \\ {}^1\!/_2 u(A) + {}^1\!/_2 u(D) &= {}^1\!/_2 u(B) + {}^1\!/_2 u(C), \\ u({}^1\!/_2 A + {}^1\!/_2 D) &= u({}^1\!/_2 B + {}^1\!/_2 C). \end{aligned}$$

Notice that now we are just comparing utilities of two distributions. We could equivalently state this entirely through the preference relation:

$$ {}^1\!/_2 A + {}^1\!/_2 D \sim {}^1\!/_2 B + {}^1\!/_2 C. $$

Furthermore, suppose we wanted to state that A is preferred to B by a multiplicative factor $\alpha > 0$ of how much C is preferred to D. Again, if we had an equivalent utility function we could write

$$\begin{aligned} u(A) - u(B) &= \alpha(u(C) - u(D)), \\ u(A) + \alpha u(D) &= u(B) + \alpha u(C), \\ \tfrac{1}{1+\alpha} u(A) + \tfrac{\alpha}{1+\alpha} u(D) &= \tfrac{1}{1+\alpha} u(B) + \tfrac{\alpha}{1+\alpha} u(C), \\ u\left(\tfrac{1}{1+\alpha} A + \tfrac{\alpha}{1+\alpha} D\right) &= u\left(\tfrac{1}{1+\alpha} B + \tfrac{\alpha}{1+\alpha} C\right), \\ \tfrac{1}{1+\alpha} A + \tfrac{\alpha}{1+\alpha} D &\sim \tfrac{1}{1+\alpha} B + \tfrac{\alpha}{1+\alpha} C. \end{aligned}$$

So we can write the same concept entirely within the preference relation.

Now if we consider a transition $t \in T$ and two distributions over histories $A, B \in \Delta(\mathcal{H})$, we can use the above to state that $t \cdot A$ is preferred to $t \cdot B$ by a multiplicative factor $\gamma(t)$ of the amount A is preferred to B:

$$\tfrac{1}{\gamma(t)+1}(t \cdot A) + \tfrac{\gamma(t)}{\gamma(t)+1} B \sim \tfrac{1}{\gamma(t)+1}(t \cdot B) + \tfrac{\gamma(t)}{\gamma(t)+1} A.$$

This brings us to what we call the Temporal γ-Indifference axiom.

Axiom 5 (Temporal γ-Indifference)**.** *For all $A, B \in \Delta(\mathcal{H})$ and transitions $t \in T$,*

$$\tfrac{1}{\gamma(t)+1}(t \cdot A) + \tfrac{\gamma(t)}{\gamma(t)+1} B \sim \tfrac{1}{\gamma(t)+1}(t \cdot B) + \tfrac{\gamma(t)}{\gamma(t)+1} A.$$

Notice this axiom is parameterized by a discount function $\gamma : T \to [0, 1]$ defined on transitions T.

The axiom essentially requires that $t \cdot A$ is preferred to $t \cdot B$ by a multiplicative factor $\gamma(t)$ of how much A is preferred to B. This fact is illustrated in the following example.

Example 1. *Suppose $\gamma(t) = 1$ for all transitions $t \in T$. Then, the temporal indifference axiom states that for all $h_1, h_2 \in \mathcal{H}$ and transitions $t \in T$,*

$$ {}^1\!/_2(t \cdot h_1) + {}^1\!/_2 h_2 \sim {}^1\!/_2(t \cdot h_2) + {}^1\!/_2 h_1. $$

In other words, given any two histories to be experienced with equal probability, the agent is indifferent to which history gets prepended with a transition, regardless of the transition. No matter which history is prepended with t, the transition t must be experienced. So the indifference is requiring the agent has no preference over which history is delayed, even if one history is highly preferred to the other.

We now state our main result. All proofs are included in the Appendix.

Theorem 4.1 (Markov Reward Theorem)**.** *A binary preference relation on $\Delta(\mathcal{H})$ satisfies Axioms 1-5 if and only if there exists a utility function $u\colon \Delta(\mathcal{H}) \to \mathbb{R}$, a reward function $r : T \to \mathbb{R}$, and transition-dependent discount function $\gamma : T \to [0, 1]$, such that $u(\varepsilon) \stackrel{\text{def}}{=} 0$, and*

$$u(t \cdot h) \stackrel{\text{def}}{=} r(t) + \gamma(t)u(h),$$

under the following conditions.

1. $\forall A, B \in \Delta(\mathcal{H})\ A \succsim B \Leftrightarrow u(A) \geq u(B)$,
2. $\forall(\sum_i p_i h_i) \in \Delta(\mathcal{H})\ u(\sum_i p_i h_i) = \sum_i p_i u(h_i)$,

where r is unique up to a positive scale factor, and γ is the function for which Axiom 5 is satisfied.

In other words, there exist a deterministic, Markov reward function such that the expected sum of rewards under a particular transition-dependent discount factor is consistent with the preference relation. Furthermore, we can show there exists an efficient algorithm that constructs the reward function and discount factor from a preference relation that satisfies Axioms 1–5 (See Algorithm 1 in the Appendix for additional details.).

Additionally, notice the form the objective takes (e.g. discounted reward, episodic total reward, average reward) is determined by the preference relation and how it satisfies the Temporal γ-Indifference axiom. This is what ultimately dictates $\gamma(t)$.

5. Objective Goals

The results thus far assume that the preferences and rewards of interest originate from the same perspective—that is, the agent doing the maximizing is the same as the agent holding the preferences. In practice, we often find ourselves in a different setting in which the relevant preferences originate from an *agent designer* that has a desired goal or purpose in mind for a separate *learning agent* to pursue (Alice and Bob in work by Abel et al. (2021), or the "designer" and "agent" in work by Singh et al. (2009)). In this section we adapt our previous assumptions to show how the results of Theorem 4.1 apply more broadly—to arbitrary sequences that contain the designer's experiences. This is what we refer to as the *objective goals* case. Indeed, this setup includes common cases from the literature where preferences are expressed in numerous ways: as demonstrations of desired behaviors (Ng et al., 2000), partial orders over a set of trajectories (Wirth et al., 2017), or through a generic interaction process with the designer (Leike et al., 2018).

In the objective goals setting, the designer experiences a stream of observations $\bar{o} \in \bar{\mathcal{O}}$. These form a distinct process which is potentially related to the observation stream of the agent. For instance, the designer may observe more than the agent: $\mathcal{O} \subset \bar{\mathcal{O}}$. In other cases, $\bar{\mathcal{O}}$ may include the agent's actions. The designer provides the agent with a learning signal that reflects its preferences on distributions over histories $\bar{h}_t = \bar{o}_1, \bar{o}_2, \ldots, \bar{o}_t$, where each $\bar{o}_i \in \bar{\mathcal{O}}$. We let $\bar{\mathcal{H}}_n$ be all histories of length n and $\bar{\mathcal{H}} \stackrel{\text{def}}{=} \bigcup_{n=0}^{\infty} \bar{\mathcal{H}}_n$ be all finite length histories over designer observations. In what follows, we suppose the designer maintains a preference relation over distributions from $\Delta(\bar{\mathcal{H}})$, then we adapt our assumptions so the results of Theorem 4.1 apply to this setting.

Assumption 5 (Objective Goals)**.** *"All of what we mean by goals and purposes" can be expressed as a binary preference relation on distributions over finite histories of designer observations $\Delta(\bar{\mathcal{H}})$.*

We have two choices here for defining the agent's interface to the environment. First, the designer can provide the rewards r and the discounts γ as separate inputs to the agent. Second, the designer can provide rewards that are already discounted,

$$r_i \stackrel{\text{def}}{=} \left(\prod_{j=1}^{i-1} \gamma(\bar{o}_j) \right) r(\bar{o}_i). \tag{2}$$

We adopt the second view to maintain the standard agent-environment interface, but note that the former view might yield an alternative plausible account.

Assumption 6 (Cumulative Sum of Objective Rewards)**.** *The "maximization of the expected value of the cumulative sum of a received scalar signal (reward)" means that there is a reward function $r : \bar{\mathcal{O}} \to \mathbb{R}$ and transition-dependent discount function $\gamma : \bar{\mathcal{O}} \to [0, 1]$, such that a designer prefers $\pi_1 \succsim_r \pi_2$ under the reward r if and only if there exists N such that $V_n^{\pi_1} > V_n^{\pi_2}$ for all $n \geq N$, where $V_n^{\pi} = E\left[\sum_{i=1}^{n} r_i\right]$.*

With this we can now restate our interpretation of the reward hypothesis for the objective goals case. What the reward hypothesis means by "well thought of" is that for any binary preference relation on distributions of a designer's histories, there exists an already discounted r, as defined in (2), such that $\pi_1 \succsim_g \pi_2$ if and only if $\pi_1 \succsim_r \pi_2$.

6. History and Related Work

We next discuss relevant literature from across the RL community. We discuss connections to the economics literature in the appendix.

6.1. Utility Theory for Sequential Decision Making

Pitis (2019) explored the relationship between the vNM axioms and the typical objectives of RL, with a focus on the

discount factor, γ. As discussed in the introduction, Pitis provides a normative account for why we should embrace a state-action dependent discount factor, as developed by White (2017): A fixed discount cannot capture all preferences we might consider rational. Second, Pitis presents three axioms on top of the vNM axioms that characterize the conditions under which a state-action dependent discount factor can be viewed as rational. These axioms bear some resemblance to aspects of our formalism: For instance, Pitis' countable transitivity axiom also addresses the issue of infinite experiences, like our Assumption 2.

Shakerinava & Ravanbakhsh (2022) build off the work of Pitis (2019) to study the existence of reward functions in a variety of controlled MDPs. In their work, Shakerinava & Ravanbakhsh formalize the notion of goals with a preference relation over outcomes generated from the given MDP.

Our work differs from Shakerinava & Ravanbakhsh (2022) in several ways. Firstly, we take into account preference relations over distributions of observation-action histories, enabling us to reason about goals in a wide range of stochastic environments, including fully-observable MDPs. Furthermore, we formalize the reward hypothesis with a set of formal assumptions. This applies to the discounted reward, average reward, and episodic total reward settings. We establish a connection between our findings and scenarios where the reward is internally generated by the agent, as well as situations involving the objective desires of an observer. We show there exists an efficient algorithm to construct rewards using a preference relation that satisfies Axioms 1–5 (See Algorithm 1 in the Appendix). Moreover, our Temporal γ-Indifference axiom provably generalizes two of their axioms. We expand on this below.

Axiom 6 (Memoryless; Shakerinava & Ravanbakhsh, 2022)**.** *For all $t \in T$ and $A, B \in \Delta(\mathcal{H})$*

$$A \succsim B \iff t \cdot A \succsim t \cdot B.$$

Theorem 6.1. *A binary preference relation on $\Delta(\mathcal{H})$ satisfies Axioms 1-5 where $\gamma(t)$ is relaxed to be in $\mathbb{R}_{\geq 0}$ if and only if Axioms 1-4 and the Memoryless axiom are satisfied.*

Axiom 7 (Additivity; Shakerinava & Ravanbakhsh, 2022)**.** *For all $h_1, h_2 \in \mathcal{H}$, $A, B, C, D \in \Delta(\mathcal{H})$ and $p \in [0, 1]$,*

$$\begin{aligned} &p(h_1 \cdot A) + (1-p)C \succsim p(h_1 \cdot B) + (1-p)D \\ \Leftrightarrow &p(h_2 \cdot A) + (1-p)C \succsim p(h_2 \cdot B) + (1-p)D. \end{aligned}$$

Theorem 6.2. *A binary preference relation on $\Delta(\mathcal{H})$ satisfies Axioms 1-5 with $\gamma(t) = 1$ if and only if Axioms 1-4 are satisfied as well as Additivity.*

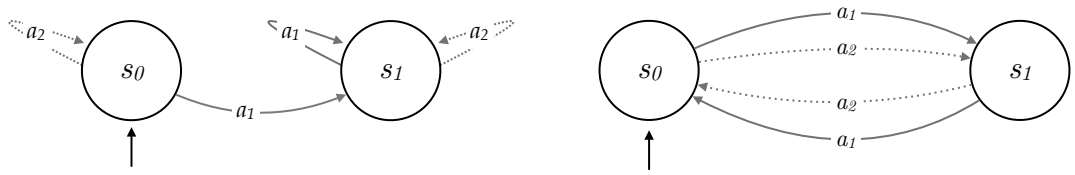
Separately, Sunehag & Hutter (2011; 2015) study what constitutes a *rational reinforcement learning agent*. More concretely, in both works, Sunehag and Hutter suppose the environment and reward function are defined, and set out to characterize what kinds of agents might we consider rational. In their first work (Sunehag & Hutter, 2011), they provide a series of properties that characterize what it means for an RL agent to be rational. These properties bear some similarity to the vNM axioms, but are agent-side properties rather than implicit requirements on the structure of goals or rewards themselves. In follow up work Sunehag & Hutter (2015) focus specifically on the rationality of optimistic agents. That is, given a reward function and environment, they contrast optimistic agents with those that are strictly expected utility maximizers. They give a full characterization of rational agency that justifies the use of optimism for exploration, showing that expected utility maximization on its own can be strictly worse than optimistic behavior.

6.2. The Limited Expressivity of Markov Reward

Abel et al. (2021) study the expressivity of reward in Markovian environments. In particular, they suppose the environment is characterized by a finite state space and transition function, and assume that preferences over objects defined with respect to the environment are given. These preferences come in three forms: (1) A set of acceptable policies, (2) A partial ordering on policies, or (3) A partial ordering on fixed-length state-action trajectories. Each of these preference types are defined with respect to the environment's *state space*, $\mathcal{S}$, that is known to be sufficient to support a Markovian transition function (and thus, $e : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$). For example, in the case of (1), the policy space is defined as all deterministic mappings of the form $\pi : \mathcal{S} \to \mathcal{A}$. Then, a choice of the first preference type is just a selection of acceptable policies. Under these three types, Abel et al. show that there are restrictions on what kinds of preferences can be codified in terms of a reward function that is Markov with respect to the same state space. Specifically, they point out two styles of counterexample: (1) the *steady state type*, in which preferences have bearing on impossible outcomes, and (2) the *entailment type*, in which the desirability of an action choice depends on behavior elsewhere in the environment.

We next show how the two styles of counterexample from Abel et al. play out in the context of our results. We find that the steady state type violates one of our assumptions, and the entailment type violates an axiom.

Steady State. First, recall the steady state counterexample, pictured in Figure 1(a). The given preference over policies asserts that the only acceptable policy chooses a_2 in the left state, but a_1 in the right. This is a counterexample in the sense that there is no Markov reward function that ensures the policy $\pi_{22} : \{s_0 \mapsto a_2 \mid s_1 \mapsto a_2\}$, has higher value than $\pi_{21} : \{s_0 \mapsto a_2 \mid s_1 \mapsto a_1\}$, since both policies never reach state s_1.



(a) Steady State Case (b) Entailment Case

Figure 1. The two counterexamples from Abel et al. (2021). In the steady state case, the set of acceptable policies contains only the policy that executes a_2 in the left, and a_1 in the right. In the entailment case, the two acceptable policies are those that choose a *different* action across the two states.

Under our formalism, these policies would induce equivalent outcome distributions and thus be interchangeably preferred. We begin with preferences over outcomes of the kind suggested by Assumption 1. That is, in the two-state environment described, our preferences are over sequences of state-action pairs. Recall that under Assumption 2, we understand a preference over policies $\pi_1 \succsim_g \pi_2$ to mean that there exists a time after which we always prefer the distribution over outcomes produced by π_1 to π_2. Thus, when we consider the steady state counter example, we can see that Assumption 2 is violated: We prefer π_1 to π_2 even though they induce *identical* distributions $D_n^{\pi_1}$ and $D_n^{\pi_2}$. In this sense, the counterexample is an instance of a broader principle we have codified in our formalism.

Entailment. The second kind of counterexample deals with cases in which the value of one policy necessarily entails something about the value of another policy. For instance, in the example pictured in Figure 1(b), the two acceptable policies are those that take a different action in each of the two states. So, $\pi_{12} : \{s_0 \mapsto a_1 \mid s_1 \mapsto a_2\}$ and $\pi_{21} : \{s_0 \mapsto a_2 \mid s_1 \mapsto a_1\}$ are both desirable because they each take opposing actions across the two states. In follow up work, Abel et al. (2022) demonstrate that a simple state-construction procedure can resolve the counterexample. In our framing, where precisely do the preferences of this entailment type go wrong? Such preferences directly violate Axiom 5, the new axiom. To see why, consider the two (Dirac) distributions: $A = s_2, a_2$ and $B = s_2, a_1$. Let $t = s_1, a_1$. That is, the composite distributions formed are $t \cdot A = s_1, a_1, s_2, a_2$ and $t \cdot B = s_1, a_1, s_2, a_1$. However, there is no choice of $\gamma(t)$ for which the indifference expressed by the Axiom holds, as the preference requires that $A \succsim B$.

7. Challenges to the Reward Hypothesis

We next summarize common challenges to the reward hypothesis and consider whether our formalization of the hypothesis provides any further insight into these arguments.

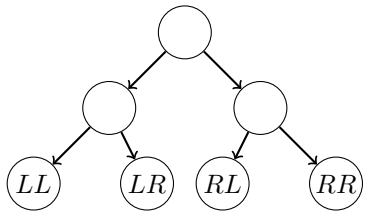
7.1. Human Irrationality

Extensive work by Kahneman & Tversky and Johnson-Laird showed how human behavior deviates from the rational model proposed by von Neumann and Morgenstern (Kahneman et al., 1982; Kahneman & Tversky, 1982; Tversky & Kahneman, 1983; Johnson-Laird, 1983). Based on these, one might conclude that our description of goals and purposes does not apply to humans, and is therefore incomplete. However, the expression of goals is distinct from the behaviors that emerge in their pursuit. Our results prove the existence of a Markov reward signal under the presumption that all goals and purposes can be rationally expressed.

7.2. Multiple Objectives

Another natural reaction to the reward hypothesis is to suspect that collapsing all of the nuance that might go into "purpose" down to a single scalar seems difficult, if not impossible. Suppose we are interested in designing an autonomous taxi to take passengers between the airport and the university. We would like the taxi to balance between safety, timeliness, and energy use. But how precisely do we make these trade-offs, and can we really reduce the nuances of their trade-offs down to a single scalar?

This challenge often yields a variant called multi-objective or multi-criteria decision making that has been studied extensively in the literature. Hausner (1953) drops continuity (Axiom 4) in order to generalize the vNM results to the multi-dimensional setting. Gábor et al. (1998) propose multi-criteria RL in MDPs and establish initial conditions for importing classical results from scalar-valued rewards such as the existence of stationary policies and the Bellman equation. More recently, Miura (2022) explicitly focus on the expressivity of multi-dimensional reward, enriching the result's of Abel et al. (2021). In particular, Miura shows that multi-dimensional Markov reward functions are strictly more expressive than scalar Markov reward functions in MDPs. However, we note that this expressivity comes with the cost of violating at least one of the axioms—we describe this in more detail shortly. Pitis et al. (2022) make a similar argument, and prove that some multi-objective problems



History	Total r_1	Total r_2
LL	1	-1
LR	0	-1
RL	1	-1
RR	0	1

Figure 2. An environment used to illustrate how Constrained MDPs can violate independence and continuity. Two actions, L and R, produce four histories shown in the table above. Also shown are the cumulative rewards of each history.

cannot be collapsed to a scalar objective.

Constrained MDPs. Constrained MDPs have been viewed as a challenge to the reward hypothesis (Szepesvári, 2020). In a Constrained MDP, the goal is to maximize the expected sum of a base scalar reward, subject to additional constraints on the expected sums of other independent reward functions. Szepesvári (2020) showed that it is generally not possible to solve these objectives as typical MDPs, with a process of "scalarization". One can show that Constrained MDPs do not always respect our notion of "goals and purposes".

Consider the example pictured in Figure 2. The environment contains two actions, L and R, whose combinations produce four different histories, denoted LL, LR, RL, RR. The cumulative payoffs under the base reward r_1 and secondary reward r_2 are also shown. Suppose the constrained objective involves maximizing expected utility under r_1 while demanding expected utility under r_2 be non-negative. Under this objective, distributions of feasible histories are preferred to those which are infeasible. Consider two such distributions $A = \frac{1}{2}RL + \frac{1}{2}RR$ and $B = LL$. We have $A \succsim B$, since the first distribution is feasible and the second is not. Independence (Axiom 3) asserts that the preference remains unchanged when A and B are equally-mixed with any other distribution. However, in this example, the preference reverses when mixed with $C = RR$. That is $\frac{1}{2}A + \frac{1}{2}C \precsim \frac{1}{2}B + \frac{1}{2}C$. In this case, both distributions are feasible, but the first achieves less expected utility under the base reward r_1.

A similar example can be used for continuity (Axiom 4). This time let $A = \frac{1}{2}RL + \frac{1}{2}RR$, $B = \frac{1}{2}LR + \frac{1}{2}RR$, and $C = RR$. We have $A \succsim B \succsim C$, because A is feasible with the highest expected base utility, B is feasible but achieves less utility than A, and C is infeasible. For this selection, there is no break even point $p \in (0, 1)$ that would make an agent indifferent between A and the mixture $pB + (1-p)C$; the latter is always infeasible.

Risk. Risk-sensitive goals provide another challenge to the reward hypothesis. Risk-sensitive objectives applied to the returns generated by a policy can, in general, cause the optimal policy to be non-Markovian (Bellemare et al., 2023). This can be readily seen in the case of maximizing a variance-penalized mean return, where the agent first experiences some uncontrolled randomness that produces two different rewards (e.g. 0 or 1) and then faces the choice of two actions with different rewards (again, 0 or 1) allowing it to choose between maximizing value or reducing variance. For sufficiently large variance penalties the optimal policy will be to choose the action leading to the opposite reward of what it had previously experienced. However, this is not possible for Markov policies.

Observe that when all optimal policies are non-Markov, such as in this example, there does not exist any Markov reward function with the same optimal policy under a risk-neutral objective. This is because the optimal policy will necessarily be Markovian. As a result, we can conclude that at least one of our assumptions or axioms must have been violated. In the above example, the issue is a violation of Axiom 5 and can be overcome by augmenting the state using the objective goals formulation.

8. Conclusion

We have here provided a conclusive account summarizing the implicit conditions required for the reward hypothesis. We separate such conditions into *assumptions*, that center around interpretations of the hypothesis, and *axioms*, that express specific formal properties about rational preference relations of all that one could mean by goals and purposes. Our main result (Theorem 4.1) states that, under Assumptions 1-4, the reward hypothesis for Markov reward functions holds if and only if Axioms 1-5 are satisfied. This result completely specifies the requirements on preferences under which the hypothesis holds. We further explore consequences of our new framing and results, including a variant from the viewpoint of the designer, an efficient constructive algorithm that translates rational preferences into a reward function, and discussion of how this axiomatic perspective can sharpen our understanding of alternative RL objectives such as constrained MDPs and risk.

9. Acknowledgements

The authors gratefully acknowledge how the interactions and feedback of their colleagues have shaped this paper. In particular, we would like to thank Andrè Barreto and Doina Precup for their thoughtful comments on an early draft of the paper, Bernardo Ávila Pires for helpful conversations, and Niko Yasui for a thorough review of the text and algorithm, including improvements to the constructive algorithm. Any remaining errors are those of the authors. The work presented here was in part supported by NSERC, CIFAR, and Amii.

References

Abel, D., Dabney, W., Harutyunyan, A., Ho, M. K., Littman, M. L., Precup, D., and Singh, S. On the expressivity of Markov reward. In *Advances in Neural Information Processing Systems*, 2021.

Abel, D., Barreto, A., Bowling, M., Dabney, W., Hansen, S., Harutyunyan, A., Ho, M. K., Kumar, R., Littman, M. L., Precup, D., and Satinder, S. Expressing non-Markov reward to a Markov agent. *Multidisciplinary Conference on Reinforcement Learning and Decision Making*, 2022.

Aumann, R. J. Utility theory without the completeness axiom. *Econometrica: Journal of the Econometric Society*, pp. 445–462, 1962.

Bellemare, M. G., Dabney, W., and Rowland, M. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

Bernoulli, D. Specimen theoriae novae de mensura sortis (trans. in 1954 as exposition of a new theory on the measurement of risk). *Econometrica*, 22(1):23–36, 1738.

Cassandra, A. R., Kaelbling, L. P., and Littman, M. L. Acting optimally in partially observable stochastic domains. In *Proceedings of the AAAI Conference on Artificiall Intelligence*, 1994.

Chang, R. Value incomparability and incommensurability. *The Oxford handbook of value theory*, pp. 205–224, 2015.

Dong, S., Van Roy, B., and Zhou, Z. Simple agent, complex environment: Efficient reinforcement learning with agent states. *arXiv preprint arXiv:2102.05261*, 2021.

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

Gábor, Z., Kalmár, Z., and Szepesvári, C. Multi-criteria reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, volume 98, 1998.

Hausner, M. Multidimensional utilities (rev). Technical report, RAND CORP SANTA MONICA CA, 1953.

Johnson-Laird, P. N. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, 1983.

Kahneman, D. and Tversky, A. The psychology of preferences. *Scientific American*, 246(1):160–173, 1982.

Kahneman, D., Slovic, S. P., Slovic, P., and Tversky, A. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

Keeney, R. L., Raiffa, H., and Meyer, R. F. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press, 1993.

Koopmans, T. C. Stationary ordinal utility and impatience. *Econometrica: Journal of the Econometric Society*, pp. 287–309, 1960.

Koopmans, T. C., Diamond, P. A., and Williamson, R. E. Stationary utility and time perspective. *Econometrica: Journal of the Econometric Society*, pp. 82–100, 1964.

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

Lattimore, T. *Theory of General Reinforcement Learning*. PhD thesis, The Australian National University, 2014.

Lattimore, T., Hutter, M., and Sunehag, P. The sample-complexity of general reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2013.

Leike, J. *Nonparametric general reinforcement learning*. PhD thesis, The Australian National University, 2016.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Lewis, A. A. On effectively computable realizations of choice functions: Dedicated to professors kenneth j. arrow and anil nerode. *Mathematical Social Sciences*, 10 (1):43–80, 1985.

Littman, M. The reward hypothesis. https://tinyurl.com/4z52r3fe, 2017.

Lu, X., Van Roy, B., Dwaracherla, V., Ibrahimi, M., Osband, I., and Wen, Z. Reinforcement learning, bit by bit. *arXiv preprint arXiv:2103.04047*, 2021.

Machina, M. J. Expected utility hypothesis. In *Utility and probability*, pp. 79–95. Springer, 1990.

Mahadevan, S. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1):159–195, 1996.

Majeed, S. J. *Abstractions of General Reinforcement Learning*. PhD thesis, The Australian National University, 2021.

Martin, R. The st. petersburg paradox. *Stanford Encyclopedia of Philosophy*, 2011.

McCarthy, J. What is artificial intelligence. *URL: http://www-formal. stanford. edu/jmc/whatisai. html*, 1998.

Miura, S. On the expressivity of multidimensional Markov reward. In *In RLDM Workshop on Reinforcement Learning as a Model of Agency*, 2022.

Ng, A. Y., Russell, S., et al. Algorithms for inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pp. 663–670, 2000.

Pitis, S. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Pitis, S., Bailey, D., and Ba, J. Rational multi-objective agents must admit non-markov reward representations. *NeurIPS Workshop on Machine Learning Safety*, 2022.

Ramsey, F. P. Truth and probability. In *Readings in formal epistemology*, pp. 21–45. Springer, 1926.

Richter, M. K. and Wong, K.-C. Computable preference and utility. *Journal of Mathematical Economics*, 32(3): 339–354, 1999.

Rustem, B. and Velupillai, K. Rationality, computability, and complexity. *Journal of Economic Dynamics and Control*, 14(2):419–432, 1990.

Shakerinava, M. and Ravanbakhsh, S. Utility theory for sequential decision making. In *Proceedings of the International Conference on Machine Learning*, 2022.

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

Singh, S., Lewis, R. L., and Barto, A. G. Where do rewards come from? In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2009.

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

Sunehag, P. and Hutter, M. Axioms for rational reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 338–352. Springer, 2011.

Sunehag, P. and Hutter, M. Rationality, optimism and guarantees in general reinforcement learning. *The Journal of Machine Learning Research*, 16(1):1345–1390, 2015.

Sutton, R. S. The reward hypothesis. http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/rewardhypothesis.html, 2004.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Szepesvári, C. Constrained MDPs and the reward hypothesis. https://readingsml.blogspot.com/2020/03/constrained-mdps-and-reward-hypothesis.html, 2020.

Tversky, A. and Kahneman, D. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293, 1983.

von Neumann, J. and Morgenstern, O. *Theory of Games and Economic Behavior*. Princeton University Press, third edition, 1953.

White, M. Unifying task specification in reinforcement learning. In *Proceedings of the Thirty-fourth International Conference on Machine Learning*, 2017.

Wirth, C., Akrour, R., Neumann, G., Fürnkranz, J., et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136): 1–46, 2017.

A. Proof of Main Theorem

Theorem 4.1 (Markov Reward Theorem)**.** *A binary preference relation on $\Delta(\mathcal{H})$ satisfies Axioms 1-5 if and only if there exists a utility function $u\colon \Delta(\mathcal{H}) \to \mathbb{R}$, a reward function $r : T \to \mathbb{R}$, and transition-dependent discount function $\gamma : T \to [0, 1]$, such that $u(\varepsilon) \stackrel{\text{def}}{=} 0$, and*

$$u(t \cdot h) \stackrel{\text{def}}{=} r(t) + \gamma(t)u(h),$$

under the following conditions.

1. $\forall A, B \in \Delta(\mathcal{H})\ A \succsim B \Leftrightarrow u(A) \geq u(B)$,
2. $\forall (\sum_i p_i h_i) \in \Delta(\mathcal{H})\ u(\sum_i p_i h_i) = \sum_i p_i u(h_i)$,

where r is unique up to a positive scale factor, and γ is the function for which Axiom 5 is satisfied.

Proof. We first show the axioms imply the existence of a Markov reward. By Theorem 3.1 and axioms 1-4, we can select $u : \Delta(\mathcal{H}) \to \mathbb{R}$, such that $u(\varepsilon) = 0$, leaving u unique up to a positive scale factor. Define $r(t) \stackrel{\text{def}}{=} u(t)$. From Axiom 5 and choosing $h_1 = h$ and $h_2 = \varepsilon$,

$$\frac{1}{\gamma(t)+1}(t \cdot h) + \frac{\gamma(t)}{\gamma(t)+1}\varepsilon \sim \frac{1}{\gamma(t)+1}(t \cdot \epsilon) + \frac{\gamma(t)}{\gamma(t)+1}h$$

Applying Theorem 3.1 (first consequence 1 then consequence 2) we get,

$$\frac{1}{\gamma(t)+1}u(t \cdot h) + \frac{\gamma(t)}{\gamma(t)+1}u(\varepsilon) = \frac{1}{\gamma(t)+1}u(t) + \frac{\gamma(t)}{\gamma(t)+1}u(h)$$

Multiplying by $\gamma(t) + 1$, we get,

$$\begin{aligned} u(t \cdot h) + \gamma(t)u(\varepsilon) &= u(t) + \gamma(t)u(h), \\ u(t \cdot h) &= r(t) + \gamma(t)u(h) \end{aligned}$$

We now show that any Markov reward satisfies the axioms. Due to Theorem 3.1 we know Axioms 1-4 are satisfied. We also know, for all $h \in \mathcal{H}$,

$$u(t \cdot h) = r(t) + \gamma(t)u(h)$$

so,

$$u(t \cdot h) - \gamma(t)u(h) = r(t).$$

Hence for all $h_1, h_2 \in \mathcal{H}$

$$u(t \cdot h_1) - \gamma(t)u(h_1) = r(t) = u(t \cdot h_2) - \gamma(t)u(h_2).$$

Rearranging we get,

$$u(t \cdot h_1) + \gamma(t)u(h_2) = u(t \cdot h_2) + \gamma(t)u(h_1).$$

Dividing all terms by $\gamma(t) + 1$,

$$\frac{1}{\gamma(t)+1}u(t \cdot h_1) + \frac{\gamma(t)}{\gamma(t)+1}u(h_2) = \frac{1}{\gamma(t)+1}u(t \cdot h_2) + \frac{\gamma(t)}{\gamma(t)+1}u(h_1).$$

Applying Theorem 3.1 (first consequence 2 then consequence 1),

$$\frac{1}{\gamma(t)+1}(t \cdot h_1) + \frac{\gamma(t)}{\gamma(t)+1}h_2 \sim \frac{1}{\gamma(t)+1}(t \cdot h_2) + \frac{\gamma(t)}{\gamma(t)+1}h_1$$

thus Axiom 5 is satisfied. □

B. Proofs of Relationship to (Shakerinava & Ravanbakhsh, 2022)

For the following proofs let $u_t(x) \stackrel{\text{def}}{=} u(t \cdot x)$ for all $t \in \mathcal{O} \times \mathcal{A}$.

Theorem 6.1. *A binary preference relation on $\Delta(\mathcal{H})$ satisfies Axioms 1-5 where $\gamma(t)$ is relaxed to be in $\mathbb{R}_{\geq 0}$ if and only if Axioms 1-4 and the Memoryless axiom are satisfied.*

Proof. Starting from the Temporal γ-Indifference Axiom, we use the vNM theorem under a specific choice of history to reduce the utility function to a positive affine form.

$\implies$ Take some t from $\mathcal{O} \times \mathcal{A}$, and let $\gamma(t) > 0$. Temporal γ-Indifference states that for any distributions A and B from $\Delta(\mathcal{H})$

$$\frac{1}{1+\gamma(t)}\,(t \cdot A) + \frac{\gamma(t)}{1+\gamma(t)}\,B \sim \frac{1}{1+\gamma(t)}\,(t \cdot B) + \frac{\gamma(t)}{1+\gamma(t)}\,A.$$

The vNM utility theorem guarantees this indifference has a utility representation:

$$\begin{aligned}\frac{1}{1+\gamma(t)}u(t \cdot A) + \frac{\gamma(t)}{1+\gamma(t)}u(B) &= \frac{1}{1+\gamma(t)}u(t \cdot B) + \frac{\gamma(t)}{1+\gamma(t)}u(A),\\ u(t \cdot A) + \gamma(t)u(B) &= u(t \cdot B) + \gamma(t)u(A).\end{aligned}$$

Let $B = \epsilon$, and define $u(\epsilon) \triangleq 0$, $r(t) \triangleq u(t \cdot \epsilon)$ so we obtain

$$u(t \cdot A) = r(t) + \gamma(t)u(A).$$

Using Lemma B.1 we get that u_t is strategically equivalent to u, where $a = r(t)$ and $b = \gamma(t)$. Hence, for all distributions $A, B \in \Delta(\mathcal{H})$

$$u(A) \geq u(B) \iff u_t(A) \geq u_t(B),$$

therefore,

$$A \succsim B \iff (t \cdot A) \succsim (t \cdot B).$$

$\impliedby$ The Memoryless Axiom states that for all $t \in T$ and $A, B \in \Delta(\mathcal{H})$

$$A \succsim B \iff (t \cdot A) \succsim (t \cdot B)$$

This means that u_t is strategically equivalent to u. By Lemma B.1, we know there exists constants, which we label $r(t)$ and $\gamma(t)$, such that,

$$u(t \cdot A) = u_t(A) = r(t) + \gamma(t)u(A), \qquad u(t \cdot B) = u_t(B) = r(t) + \gamma(t)u(B).$$

Simple algebra shows that

$$\begin{aligned}u(t \cdot A) - \gamma(t)u(A) &= u(t \cdot B) - \gamma(t)u(B),\\ u(t \cdot A) + \gamma(t)u(B) &= u(t \cdot B) + \gamma(t)u(A)\\ \frac{1}{1+\gamma(t)}u(t \cdot A) + \frac{\gamma(t)}{1+\gamma(t)}u(B) &= \frac{1}{1+\gamma(t)}u(t \cdot B) + \frac{\gamma(t)}{1+\gamma(t)}u(A),\\ \frac{1}{1+\gamma(t)}(t \cdot A) + \frac{\gamma(t)}{1+\gamma(t)}B &\sim \frac{1}{1+\gamma(t)}(t \cdot B) + \frac{\gamma(t)}{1+\gamma(t)}A.\end{aligned}$$

The last line follows from the vNM utility theorem. □

Definition 1 (Strategic Equivalence)**.** *Two utility functions u_1 and u_2 are strategically equivalent if and only if they imply the same preference ranking for any two distributions.*

Lemma B.1. *Two utility functions u_1 and u_2 are strategically equivalent if and only if there exists two constants a and $b > 0$ such that*

$$u_1(x) = a + bu_2(x) \text{ for all } x.$$

Proof. The first direction follows a similar argument to Theorem 4.1 from Keeney et al. (1993).

$\implies$ Let $x_0 \in \arg\min_x u_2(x)$ and $x^* \in \arg\max_x u_2(x)$. We first consider the degenerate case where $u_2(x_0) = u_2(x^*)$, so u_2 is a constant function. By strategic equivalence u_1 must also be a constant function, trivially satisfying the implication. Now consider the case where $u_2(x^*) > u_2(x_0)$. For any x, there exists $c \in [0, 1]$ such that $x \sim cx^* + (1 - c)x_0$ under u_2, and by strategic equivalence u_1 as well. Therefore,

$$u_i(x) = cu_i(x^*) + (1 - c)u_i(x_0) \text{ for } i = 1, 2.$$

Letting $i = 2$ and solving for c we get

$$c = \frac{u_2(x) - u_2(x_0)}{u_2(x^*) - u_2(x_0)}.$$

Substituting this value of c in when $i = 1$ gives.

$$\begin{aligned} u_1(x) &= \left(\frac{u_2(x) - u_2(x_0)}{u_2(x^*) - u_2(x_0)}\right) u_1(x^*) + \left(1 - \frac{u_2(x) - u_2(x_0)}{u_2(x^*) - u_2(x_0)}\right) u_1(x_0), \\ &= \underbrace{\left(u_1(x_0) + \frac{u_2(x_0)u_1(x_0) - u_2(x_0)u_1(x^*)}{u_2(x^*) - u_2(x_0)}\right)}_{a} + \underbrace{\left(\frac{u_1(x^*) - u_1(x_0)}{u_2(x^*) - u_2(x_0)}\right)}_{b} u_2(x). \end{aligned}$$

Notice that $b > 0$, because $u_2(x^*) > u_2(x_0)$.

$\impliedby$ Assume there exists constants a and $b > 0$ such that

$$u_1(x) = a + bu_2(x) \text{ for all } x.$$

Take any two x and x' such that $u_2(x) > u_2(x')$. Scaling by a positive constant b and shifting the utility by a leaves the relation unchanged. Therefore,

$$\begin{aligned} u_2(x) &> u_2(x'), \\ a + bu_2(x) &> a + bu_2(x'), \\ u_1(x) &> u_1(x'). \end{aligned}$$

□

Theorem 6.2. *A binary preference relation on $\Delta(\mathcal{H})$ satisfies Axioms 1-5 with $\gamma(t) = 1$ if and only if Axioms 1-4 are satisfied as well as Additivity.*

Proof. The first direction follows from Independence and Lemma B.2. The converse is reduced to the Memoryless Axiom, then the remainder of the argument follows from Theorem 6.1.

$\implies$ Take any $t \in \mathcal{O} \times \mathcal{A}$, and $A, B, C, D \in \Delta(\mathcal{H})$ for which

$$p(t \cdot A) + (1 - p)C \succsim p(t \cdot B) + (1 - p)D.$$

By the Independence Axiom, the preference remains unchanged after mixing distributions with $(t' \cdot X)$ by an amount $q \in [0, 1]$:

$$qp(t \cdot A) + q(1 - p)C + (1 - q)(t' \cdot X) \succsim qp(t \cdot B) + q(1 - p)D + (1 - q)(t' \cdot X).$$

If we take $qp = (1-q)$, then we can form uniform compound distributions of $(t, \cdot)$ and $(t' \cdot X)$:

$$\begin{aligned} qp[(t \cdot A) + (t' \cdot X)] + q(1-p)C &\succsim qp[(t \cdot B) + (t' \cdot X)] + q(1-p)D, \\ 2qp\left[\frac{1}{2}(t \cdot A) + \frac{1}{2}(t' \cdot X)\right] + q(1-p)C &\succsim 2qp\left[\frac{1}{2}(t \cdot B) + \frac{1}{2}(t' \cdot X)\right] + q(1-p)D, \\ 2qp\left[\frac{1}{2}(t' \cdot A) + \frac{1}{2}(t \cdot X)\right] + q(1-p)C &\succsim 2qp\left[\frac{1}{2}(t', B) + \frac{1}{2}(t \cdot X)\right] + q(1-p)D. \end{aligned}$$

The last line follows from Lemma B.2, which takes Temporal Indifference as its precondition.

Finally, we invoke the Independence axiom to remove $(t \cdot X)$ mixture:

$$p(t' \cdot A) + (1-p)C \succsim p(t', B) + (1-p)D.$$

$\Longleftarrow$ For any $t, t' \in T$, and $A, B, C, D \in \Delta(\mathcal{H})$, the Additivity Axiom states

$$p(t, A) + (1-p)C \succsim p(t \cdot B) + (1-p)D \iff p(t', A) + (1-p)C \succsim p(t', B) + (1-p)D.$$

This reduces to the Memoryless Axiom if we restrict t to $\mathcal{O} \times \mathcal{A}$, and take $t' = \epsilon$ and $C = D$:

$$\begin{aligned} p(t, A) + (1-p)D \succsim p(t \cdot B) + (1-p)D &\iff p(\epsilon, A) + (1-p)D \succsim p(\epsilon, B) + (1-p)D, \\ p(t, A) + (1-p)D \succsim p(t \cdot B) + (1-p)D &\iff pA + (1-p)D \succsim pB + (1-p)D, \\ (t, A) \succsim (t \cdot B) &\iff A \succsim B. \end{aligned}$$

The last line follows from independence on D.

Finally, Theorem 6.1 established that the Memoryless Axiom holds if and only if the Temporal γ-Indifference Axiom holds. Therefore, the remainder of the proof follows from that. □

Lemma B.2. *If Temporal γ-Indifference holds when $\gamma = 1$, along with Axioms 1-5, then for any $t, t' \in \mathcal{O} \times \mathcal{A}$, and $A, X \in \Delta(\mathcal{H})$,*

$$\frac{1}{2}(t \cdot A) + \frac{1}{2}(t' \cdot X) \sim \frac{1}{2}(t' \cdot A) + \frac{1}{2}(t \cdot X).$$

Proof. Take some t from $\mathcal{O} \times \mathcal{A}$, and let $\gamma(t) = 1$. Temporal γ-Indifference states that for any distributions A and X

$$\frac{1}{2}(t \cdot A) + \frac{1}{2}X \sim \frac{1}{2}(t \cdot X) + \frac{1}{2}A.$$

This applies to any t, so it must apply to any other t' from $\mathcal{O} \times \mathcal{A}$ with $\gamma(t') = 1$:

$$\frac{1}{2}(t \cdot A) + \frac{1}{2}X \sim \frac{1}{2}(t \cdot X) + \frac{1}{2}A \iff \frac{1}{2}(t' \cdot A) + \frac{1}{2}X \sim \frac{1}{2}(t' \cdot X) + \frac{1}{2}A.$$

The vNM utility theorem guarantees these preferences have a utility representation, meaning:

$$\frac{1}{2}u(t \cdot A) + \frac{1}{2}u(X) = \frac{1}{2}u(t \cdot X) + \frac{1}{2}u(A) \iff \frac{1}{2}u(t' \cdot A) + \frac{1}{2}u(X) = \frac{1}{2}u(t' \cdot X) + \frac{1}{2}u(A).$$

Regardless of whether a distribution involves t or t', the difference between utilities will be equal:

$$\begin{aligned} u(t \cdot A) + u(X) - u(t \cdot X) - u(A) &= u(t' \cdot A) + u(X) - u(t' \cdot X) - u(A), \\ u(t \cdot A) - u(t \cdot X) &= u(t' \cdot A) - u(t' \cdot X), \\ u(t \cdot A) + u(t' \cdot X) &= u(t' \cdot A) + u(t \cdot X), \\ u\left(\frac{1}{2}(t \cdot A) + \frac{1}{2}(t' \cdot X)\right) &= u\left(\frac{1}{2}(t' \cdot A) + \frac{1}{2}(t \cdot X)\right), \\ \frac{1}{2}(t \cdot A) + \frac{1}{2}(t' \cdot X) &\sim \frac{1}{2}(t' \cdot A) + \frac{1}{2}(t \cdot X). \end{aligned}$$

□

C. Preferences in the Average Reward Setting

The notion of the cumulative sum eventually being larger allows us to capture settings where the series is convergent (i.e., $\lim_{n\to\infty} V_n^\pi$ exists) as well as cases where it is not, such as average reward; if one policy has higher average reward, then its finite sum must eventually be larger. For this analysis we will assume without loss of generality that the transition-dependent discount $\gamma(t) = 1\ \forall t \in T$.

This last point is made formal in the following proposition, where the average reward for policy π is $\mu_\pi \triangleq \lim_{n\to\infty} \frac{1}{n} V_n^\pi$.

Proposition C.1. *For any policies π_A, π_B whose average rewards μ_A, μ_B exist, if $\mu_A > \mu_B$, then there exists an N such that $V_n^A > V_n^B$ for all $n \geq N$.*

Proof. Assume $\mu_A > \mu_B$ exist. According to the definition of a limit, for any $\varepsilon_A, \varepsilon_B > 0$, there exists some N_A and N_B such that

$$\mu_A - \varepsilon_A < \frac{1}{k} V_k^A \qquad\qquad \mu_B + \varepsilon_B > \frac{1}{m} V_m^B,$$

for all $k \geq N_A$ and $m \geq N_B$.

Choose $\varepsilon_A = \varepsilon_B = \frac{1}{2}(\mu_A - \mu_B) > 0$ and let $N = \max\{N_A, N_B\}$, so the above inequalities hold for all $n \geq N$. If you negate the second inequality and add them together, then for all $n > N$,

$$\begin{aligned}
(\mu_A - \varepsilon_A) - (\mu_B + \varepsilon_B) &< \frac{1}{n} V_n^A - \frac{1}{n} V_n^B \\
(\mu_A - \mu_B) - (\varepsilon_A + \varepsilon_B) &< \frac{1}{n}\left(V_n^A - V_n^B\right) \\
(\mu_A - \mu_B) - (\mu_A - \mu_B) &< \frac{1}{n}\left(V_n^A - V_n^B\right) \\
0 &< \frac{1}{n}\left(V_n^A - V_n^B\right)
\end{aligned}$$

Thus, $V_n^A > V_n^B$. □

We can show a similar result for the specialized bias-optimal case of average reward.

Definition 2 (Bias Value)**.** *The relative difference in total reward gathered is*

$$V^\pi \triangleq \lim_{n\to\infty} \mathbf{E}\left[\sum_{i=1}^{n} (R_i^\pi - \mu_\pi)\right].$$

Proposition C.2. *For any policies π_A, π_B whose average rewards μ_A, μ_B exist, if $\mu_A = \mu_B$ and the bias values exist with $V^A > V^B$, then there exists an N such that $V_n^A > V_n^B$ for all $n \geq N$.*

Proof. Assume $\mu_A = \mu_B$ exist as well as the bias optimal values $V^A > V^B$.

$$V^A \triangleq \lim_{n\to\infty} \mathbf{E}\left[\sum_{i=1}^{n} (R_i^A - \mu_A)\right], \qquad\qquad V^B \triangleq \lim_{n\to\infty} \mathbf{E}\left[\sum_{i=1}^{n} (R_i^B - \mu_B)\right].$$

Distributing the expectation, breaking the summation, and recognizing the expected n-step sums as $V_n^A = \mathbf{E}[\sum_{i=1}^n R_i^A]$ and $V_n^B = \mathbf{E}[\sum_{i=1}^n R_i^B]$, we have

$$V^A = \lim_{n\to\infty} (V_n^A - n\mu_A), \qquad\qquad V^B \triangleq \lim_{n\to\infty} (V_n^B - n\mu_B).$$

According to the definition of a limit, for any $\varepsilon_A, \varepsilon_B > 0$, there exists some N_A and N_B such that

$$V^A - \varepsilon_A < V_k^A - k\mu_A \qquad\qquad V^B + \varepsilon_B > V_m^B - m\mu_B,$$

for all $k \geq N_A$ and $m \geq N_B$.

Choose $\varepsilon_A = \varepsilon_B = \frac{1}{2}(V^A - V^B) > 0$ and let $N = \max\{N_A, N_B\}$, so that for all $n \geq N$, the above inequalities hold. If you negate the second inequality and add them together, then for all $n > N$,

$$(V^A - \varepsilon_A) - (V^B + \varepsilon_B) < \frac{1}{n}V_n^A - \frac{1}{n}V_n^B - n\mu_A + n\mu_B$$

Since $\mu_A = \mu_B$,

$$\begin{aligned}
(V^A - \varepsilon_A) - (V^B + \varepsilon_B) &< \frac{1}{n}\left(V_n^A - V_n^B\right) \\
(V^A - V^B) - (\varepsilon_A + \varepsilon_B) &< \frac{1}{n}\left(V_n^A - V_n^B\right) \\
(V^A - V^B) - (V^A - V^B) &< \frac{1}{n}\left(V_n^A - V_n^B\right) \\
0 &< \frac{1}{n}\left(V_n^A - V_n^B\right)
\end{aligned}$$

Thus, $V_n^A > V_n^B$. □

D. A Constructive Algorithm

We now develop an algorithm that can construct the realizing reward function given a preference relation that is known to satisfy Axioms 1-5. Algorithm 1 uses the preference relation to sort outcomes, then computes rewards and two-step utilities by scaling their rankings relative to their break-even point with the best and worst outcomes. With this information, the discount factor can be computed in closed-form. Below we summarize the procedures for using a preference relation to sort and scale outcomes.

PrefSort (Algorithm 2) is a procedure for sorting a set of outcomes according to their preference. Our implementation takes in a preference relation and set of outcomes $\mathcal{T}$. The procedure returns a tuple of outcomes which are sorted in ascending order, according to $\mathcal{R}$, with MergeSort.

PrefScale (Algorithm 3) is a procedure to determine the relative degree of preference between outcomes. In our implementation, it takes a preference relation, a tuple of preference-sorted outcomes $\mathcal{T}$, and a tolerance parameter $\varepsilon \in (0, 1]$. The procedure returns a set of numerical scale factors that reflect the degree to which each outcome is preferred relative to best and worst outcomes. Our implementation assigns scale factors using a binary line search informed by the continuity axiom Axiom 4. The inner loop of the line search terminates when the difference between subsequent factors differ less than a pre-specified ε.

The complexity of Algorithm 1 only depends on $|\mathcal{A}|$ and $|\mathcal{O}|$. The call to PrefSort requires $O(n \log n)$ operations, where $n = 2|\mathcal{A} \times \mathcal{O}|$, ignoring an additive constant. The call to PrefScale requires $O(n)$ operations. We then run two for loops, the largest of which iterating through $|\mathcal{U}| = 2|\mathcal{A} \times \mathcal{O}| + 2$ elements. Thus, the total run-time is $O\left(2|\mathcal{A} \times \mathcal{O}| \log |\mathcal{A} \times \mathcal{O}|\right)$.

E. Additional Comments on Objective Goals

As a special case, consider when the environment can be modeled as a Partially Observable MDP (POMDP, (Cassandra et al., 1994)), and suppose that the designer observes the environment states and the agent's actions, $\bar{\mathcal{O}} = \mathcal{S} \times \mathcal{A}$, so that histories are $\bar{h} = a_1, s_1, a_2, s_2, \ldots$. If Axioms 1-5 apply to the designer's preference relation on this distributions over histories, where the reference to transition $(\mathcal{O} \times \mathcal{A})$ in Axiom 5 now refers to $(\mathcal{S} \times \mathcal{A})$, then Theorem 4.1 extends to reward functions $r : (\mathcal{S} \times \mathcal{A}) \to \mathbb{R}$ and discount functions $\gamma : (\mathcal{S} \times \mathcal{A}) \to [0, 1]$.

Note that the in the objective goals setting, the designer's states may encode more than necessary to produce Markov state transitions. For example, they could encode a reward bundle (Abel et al., 2022), a finite state machine defining an otherwise non-Markov reward. This should allow us to define a variant of Axiom 5 that only needs to be satisfied when state transitions consist of POMDP states produced with any other finite state machine that transitions on POMDP transitions. In other words, preferences that can be expressed as reward bundles are captured by this extended axiom.

Algorithm 1 Reward and Discount Design

INPUT: $\mathcal{R} = \{\prec, \precsim, \succ, \succsim, \sim\}$.
OUTPUT: $r : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$, $\gamma : \mathcal{A} \times \mathcal{O} \to [0, 1]$.

$\mathcal{T}_1 = \mathcal{A} \times \mathcal{O} \cup \{\varepsilon\}$
$\mathcal{T}_2 = \{t \cdot t : t \in \mathcal{T}_1\}$
$\mathcal{U} = \text{PrefSort}(\mathcal{R}, \mathcal{T}_1 \cup \mathcal{T}_2)$
$\mathcal{P} = \text{PrefScale}(\mathcal{R}, \mathcal{U}, \epsilon)$
Let i_ε be the index of ε in $\mathcal{U}$.
for $i \in \{1, \cdots, |\mathcal{U}|\}$ **do**
  $u(\tau_i) = p_{i_\varepsilon} - p_i$
end for
for $i \in \{1, \cdots, |\mathcal{T}_1|\}$ **do**
  $r(t_i) = u(t_i)$
  $\gamma(t_i) = u(t_i \cdot t_i)/u(t_i) - 1$
end for
return r, γ

Algorithm 2 PrefSort

INPUT: $\mathcal{R}, \mathcal{T} = (t_1, \cdots, t_n)$.
OUTPUT: Sorted $\mathcal{T}$.

if $n \leq 1$ **then**
  output $\mathcal{T}$
end if
$\mathcal{L}, \mathcal{K} \leftarrow ()$
for $i = 1, \cdots, n$ **do**
  if $t_i \succsim t_{\lfloor n/2 \rfloor}$ **then**
    $\mathcal{K} \leftarrow \text{concat}(\mathcal{K}, t_i)$
  else
    $\mathcal{L} \leftarrow \text{concat}(\mathcal{L}, t_i)$
  end if
end for
$\mathcal{L} \leftarrow$PrefSort$(\mathcal{L})$
$\mathcal{K} \leftarrow$PrefSort$(\mathcal{K})$
$\mathcal{T} \leftarrow \text{concat}(\mathcal{L}, \mathcal{K})$
output $\mathcal{T}$

F. Additional Comments on Multiple Objectives

Here we examine the temporal nature of handling multiple objectives in view of an agent's lived experience. We present an additional axiom with which we may want to constrain goals and purposes. Let $A[h \to B]$ refer to the distribution over histories where all histories with non-zero support in A that share the prefix h are removed, and their support shifted to $h \cdot B$.

Axiom 8 (Sequential Consistency)**.** *For all $A, B, C \in \Delta(\mathcal{H})$ and $h \in \mathcal{H}$, $h \cdot A \succ h \cdot B$ if and only if $C[h \to A] \succ C[h \to B]$.*

In other words, extensions to a hypothetical history that are more aligned with some goal or purpose do not change if that hypothetical history becomes certain. In behavioral terms, goal-aligning behavior following a possible future is the same as goal-aligning behavior following that future occurring. Alternatively, goal-aligning behavior after some history should not depend on hypothetical alternative pasts that did not come about.

This is related to notions of *dynamic inconsistency* in behavioural economics[5]: one may prefer to receive \$110 in 101 days to \$100 in 100 days, and yet when 100 days passes the same person may now prefer \$100 to waiting one day for \$110 (i.e.,

[5]It is best to think of "time passing" in our formalization as "the past" at some time t becoming certain — fixing $h_t \in \mathcal{H}_t$ — and so the policy and environment specify distributions over histories whose support has h_t as a prefix.

Algorithm 3 PrefScale

INPUT: $\mathcal{R}, \mathcal{T} = (t_1 \precsim t_2 \precsim \cdots), \epsilon \in (0, 1]$.
OUTPUT: $\mathcal{P} = \{p_1, \cdots, p_{|\mathcal{T}|}\}$.

for $i \in \{1, \cdots, |\mathcal{T}|\}$ **do**
  $p_i^{k-1} \leftarrow 1$
  $\Delta \leftarrow 2\epsilon$
  while $\Delta \geq \epsilon$ **do**
    if $t_i \succ p_i^{k-1} t_1 + (1 - p_i^{k-1}) t_{|\mathcal{T}|}$ **then**
      $p_i^k \leftarrow 3p_i^{k-1}/2$
    else if $t_i \prec p_i^{k-1} t_1 + (1 - p_i^{k-1}) t_{|\mathcal{T}|}$ **then**
      $p_i^k \leftarrow p_i^{k-1}/2$
    else
      break
    end if
    $\Delta \leftarrow |p_i^k - p_i^{k-1}|$
    $p_i^{k-1} \leftarrow p_i^k$
  end while
  $p_i = p_i^{k-1}$
end for
output $\mathcal{P}$.

human preferences are often not dynamically consistent). It is straightforward to see that Independence (Axiom 3) implies Sequential Consistency (Axiom 8). However, while one may reject Independence, it seems much more difficult to reject Sequential Consistency, i.e., the goal-alignment of extensions of histories change if the history becomes certain. Notice, however, that Constrained MDP formulations for capturing multiple objectives, in fact, violate this more specific axiom.

G. Additional Related Work

G.1. Economics

Economics has been studying the nature of rational behavior for centuries. In the 1700s, Gabriel Cramer and Daniel Bernoulli independently formulated what is now called the Expected Utility Hypothesis (Machina, 1990) in response to the "St. Petersburg Paradox" (Martin, 2011) articulated by Daniel's cousin Nicholas (Bernoulli, 1738). The Expected Utility Hypothesis says, roughly, that individuals "might maximize the expectation of 'utility' rather than of monetary value." (Machina, 1990). Centuries later, Ramsey (1926) provided the first formal axiomatic treatment of expected utility, which would later be refined by von Neumann & Morgenstern (1953) to form the now widely adopted foundations of decision theory. Following this development, an expansive body of research has explored how to account for other aspects of rationality, including uncertainty (Kreps & Porteus, 1978), time (Koopmans, 1960; Koopmans et al., 1964), and computation (Lewis, 1985; Rustem & Velupillai, 1990; Richter & Wong, 1999).

H. Constant Discounting

In this section, we comment on the constant discounting case, which is commonly employed in practice. According to our results, if the preference relation satisfies Temporal γ-Indifference with a constant discount, then there exists a Markov reward that can effectively express the desired goal. However, it's important to note that Pitis (2019) highlights the limitation of using a constant discount to express general preferences. This observation holds true, particularly in episodic problems where the discount is set to zero upon termination and remains at a constant positive value less than one elsewhere. While our results encompass scenarios where the discount remains constant and is applied exponentially, we do not specifically address situations where the discount is applied in a hyperbolic manner, as discussed by Fedus et al. (2019).