From preferences to rewards: A brief introduction

Fernando E. Rosas

April 14, 2026

Abstract

This note develops a short route from preferences over complete trajectories to expected utility, reward, and discount. We begin with preference relations on deterministic trajectories and explain how completeness and transitivity yield an ordinal utility representation. We then show how lotteries, together with the von Neumann–Morgenstern axioms, produce a cardinal utility over trajectories, and we clarify the distinction between ordinal preference utility and vNM utility. Next, following Bowling et al., we add a fifth temporal axiom that is necessary and sufficient for a recursive representation in terms of local rewards and discounting. Finally, we explain why reward is not unique: different reward functions can encode the same utility or the same preference ordering, affine changes of utility induce corresponding changes of reward, and potential-based shaping provides a canonical example of reward equivalence.

1 Introduction

What does it mean for a system to have a goal? In sequential decision making, the object of evaluation is often not a single isolated choice or prize but a whole trajectory: a complete history of states, observations, actions, or consequences unfolding through time. Before asking how goals are achieved, it is worth asking a more basic question: what does it mean to prefer some trajectories over others, and what follows from that?

Recent work in reinforcement learning has revived this fundamental question, treating preferences over histories as primitive and asking what additional assumptions are needed before one can recover scalar reward functions. The present note gives a short introduction to that path, which proceeds in three stages.

(1) First, one asks for a numerical representation of how whole trajectories are ranked.

(2) Second, once one allows lotteries over trajectories, one asks when these lotteries can be ranked by the expectation of that trajectory utility.

(3) Third, one asks what additional requirements are needed in order to decompose this expected utility into rewards assigned at each time step.

The first and second steps are closely related to the classical theory developed by von Neumann and Morgenstern (von Neumann and Morgenstern, 1944). The third asks what extra temporal structure is needed before utility over whole trajectories can be decomposed into stagewise rewards, following the line of work developed in modern reinforcement learning by Pitis (2019), Shakerinava and Ravanbakhsh (2022), and Bowling et al. (2023).

2 Preferences over trajectories

Following Bowling et al. (2023), it is useful to define trajectories in terms of the primitive interaction alphabet. Let $\mathcal{O}$ be a finite set of observations and $\mathcal{A}$ a finite set of actions. A one-step interaction is then given by $t = (o, a) \in \mathcal{O} \times \mathcal{A}$. For each $n \in \mathbb{N}_{\geq 0}$, define the space of trajectories of length n by

$$\mathcal{H}_n := (\mathcal{O} \times \mathcal{A})^n.$$

We write ε for the unique trajectory of length 0. The space of all *finite* trajectories is

$$\mathcal{H}^* := \bigcup_{n=0}^{\infty} \mathcal{H}_n.$$

A typical element of $\mathcal{H}^*$ has the form $h = (o_1, a_1, o_2, a_2, \ldots, o_n, a_n)$. We will keep the notation $\mathcal{H}^*$ for the set of all finite trajectories throughout.

Definition 2.1 (Weak preference)**.** *A weak preference relation on $\mathcal{H}^*$ is a binary relation $\succcurlyeq$ where*

$$h \succcurlyeq h'$$

means that trajectory h is judged at least as good as trajectory h'.

From $\succcurlyeq$ we derive the usual companion relations:

$$h \sim h' \iff h \succcurlyeq h' \text{ and } h' \succcurlyeq h, \qquad h \succ h' \iff h \succcurlyeq h' \text{ and not } h' \succcurlyeq h.$$

We call $\sim$ 'indifference', as an agent has no reason to prefer one over the other.

At this point, $\succcurlyeq$ has no properties whatsoever. One may naturally wonder what kinds of properties it is reasonable to require of $\succcurlyeq$, and what follows from them — which is what we study in the next sections.

3 When are preferences problematic?

Can a preference relation be intrinsically 'bad'? The relevant concern here is whether it leads to some form of self-defeat, avoidable loss, or failure of coherent guidance. The discussion in the decision-theory literature suggests at least three grades of concern.

1. *Representability failures.* A first and weakest concern is that a preference relation may fail to be representable in a convenient form. This does not, by itself, imply that the preference is irrational. Representation failures may merely be inconvenient, but they become more significant when they are symptoms of deeper issues of the kinds described next (Aumann, 1962; Fishburn, 1970).

2. *Self-defeat and avoidable loss.* A more serious concern is that preferences may guide choice poorly — as judged by the agent's own view. One important case is *static self-defeat*: choosing an option that is worse than another available one, or adopting a policy that is systematically improvable. A classic case of suboptimality is dominance: one option or policy dominates another when it is at least as good in every relevant respect and strictly better in some, so choosing the dominated option is a clear mistake (Kreps, 1988; Mas-Colell et al., 1995). A second case is *diachronic self-defeat*: a plan that the agent endorses now is predictably undone later in a way that leaves the agent worse off overall.

3. *Vulnerability.* The most vivid coherence arguments show that a collection of individually acceptable choices can be combined into a guaranteed loss. Dutch-book arguments play this role for credences; money-pump arguments play the analogous role for preferences. The standard example is a preference cycle

$$h_1 \succ h_2, \qquad h_2 \succ h_3, \qquad h_3 \succ h_1.$$

If the agent is willing to pay a small fee to move each time to a strictly preferred trajectory, then an adversary can guide it around the cycle and back to where it started, poorer than before (Gustafsson, 2010). This is why intransitivity is usually regarded as a particularly severe pathology: it is not merely hard to represent, but also vulnerable to exploitation under natural trading assumptions.

Coherence and selection.

It is useful to separate *coherence* arguments from *selection-theorem* style arguments. A coherence argument is a within-agent claim: it says that if a single agent violates some structural constraint, then the agent is vulnerable to a penalty such as a money pump, Dutch book, or dominated choice. A selection argument is different. It asks whether agents lacking that property would tend to disappear under some broader optimization pressure, such as market competition, training dynamics, or evolutionary selection.

Coherence and selection arguments are related, but not identical. A coherence result may help motivate a selection story, yet it does not by itself show that realistic environments actually select against the offending preference pattern. Conversely, a representational failure with no plausible selection story may still be mathematically interesting while being less central for explaining the structure of real agents.

For the purposes of this note, the key point is that these notions of badness do not all coincide. A preference can fail standard representation without being exploitably incoherent, and not every departure from expected utility is thereby a disaster. Nevertheless, the axioms studied below are useful because — as we will see — they rule out several important undesirable properties.

It is also helpful to distinguish three kinds of formal results. A *representation result* says that if preferences satisfy certain axioms, then they can be written in a particular mathematical form; the vNM and Savage theorems are the classical examples. A *coherence result* is different: it links violations of some constraint to a penalty such as a Dutch book, money pump, dynamic inconsistency, or dominated choice. Complete-class and admissibility theorems belong more naturally in this second family than in the first, since they characterize undominated decision rules rather than utility representations. A *selection result* is different again: it adds a story about some optimization process — such as market competition, evolution, or training dynamics — and argues that systems lacking a certain property tend to be selected against. In short, representation concerns *form*, coherence concerns *vulnerability or domination*, and selection concerns *survival under pressure*.

4 Completeness, transitivity, and ordinal utility

Two structural conditions are especially important.

Definition 4.1 (Completeness)**.** *A preference relation $\succcurlyeq$ on $\mathcal{H}^*$ is* complete *if for every pair $h, h' \in \mathcal{H}^*$,*

$$h \succcurlyeq h' \quad \text{or} \quad h' \succcurlyeq h \quad \text{(or both)}.$$

Definition 4.2 (Transitivity)**.** *A preference relation $\succcurlyeq$ on $\mathcal{H}^*$ is* transitive *if for every $h, h', h'' \in \mathcal{H}^*$,*

$$h \succcurlyeq h' \quad \text{and} \quad h' \succcurlyeq h'' \quad \Longrightarrow \quad h \succcurlyeq h''.$$

Completeness says that the agent can compare any two trajectories. This is a claim about *comparability*, not about certainty: it says that the preference relation returns a verdict on every pair, not that the agent knows everything about the consequences of those trajectories. Transitivity says that these verdicts fit together consistently across chains of comparison. For example, if $h_1 \succcurlyeq h_2$ and $h_2 \succcurlyeq h_3$, then transitivity requires $h_1 \succcurlyeq h_3$ as well.

A preference relation satisfying both conditions is often called a *weak order* in economics and decision theory (Fishburn, 1970; Kreps, 1988; Mas-Colell et al., 1995). On a countable domain such as $\mathcal{H}^*$, weak orders admit an ordinal utility representation. More general representation theorems on richer domains go back to the classic work of Debreu (1954).

Proposition 4.3 (Ordinal representation on the trajectory space)**.** *A preference relation $\succcurlyeq$ on $\mathcal{H}^*$ is complete and transitive if and only if there exists a function $u\colon \mathcal{H}^* \to \mathbb{R}$ such that*

$$h \succcurlyeq h' \iff u(h) \geq u(h') \qquad \text{for all } h, h' \in \mathcal{H}^*.$$

Proof. If such a function u exists, then completeness and transitivity are inherited from the total order on $\mathbb{R}$. Conversely, if $\succcurlyeq$ is complete and transitive, then trajectories can be grouped into indifference classes, where each class contains all trajectories tied with one another. The quotient $\mathcal{H}^*/\!\sim$ of these classes is a countable total order. Any countable total order can be embedded in $\mathbb{R}$, so we may assign real numbers to the indifference classes in a way that preserves their order. Composing that assignment with the quotient map gives the desired utility function on trajectories. □

This utility is *ordinal*: only the ranking matters. Any strictly increasing transformation of u represents the same preference relation. So ordinal utility lets us encode the order of deterministic trajectories, but it does not yet tell us how to compare risky mixtures of them. The numbers themselves have no independent meaning beyond the order they induce. If u represents a preference relation, then so does $2u + 17$, or $\exp(u)$, provided the transformation remains strictly increasing. This is why ordinal utility is best thought of as a numerical *labeling of ranks*, not yet as a measure of how much one trajectory is preferred to another (Fishburn, 1970).

It is also useful to understand what happens when either condition fails.

What if completeness or transitivity fail

If completeness fails. Then some pairs of trajectories are incomparable. This may be a feature rather than a bug: the agent may genuinely refuse to rank certain alternatives because its values are plural, under-specified, or context-sensitive. But once incomparability is allowed, no single real-valued utility function can exactly represent the relation in the sense of Proposition 4.3, because any two real numbers are themselves comparable. One must then move to a different formalism, such as partial orders, sets of utility functions, or multi-criteria representations (Aumann, 1962).

If transitivity fails. Then local pairwise judgments need not assemble into a global ranking. In the simplest case one obtains a cycle

$$h_1 \succ h_2, \qquad h_2 \succ h_3, \qquad h_3 \succ h_1.$$

No scalar utility can represent such a cycle, since it would require

$$u(h_1) > u(h_2) > u(h_3) > u(h_1),$$

which is impossible. The logical problem is already serious: there is no single global ranking of the options. Under the additional behavioral assumption that the agent is willing to pay a small fee to move from any option to a strictly preferred one, this logical failure becomes operational through the *money pump* problem. Suppose an agent currently has h_1 and is willing to pay a small fee $\varepsilon > 0$ each time it moves to a strictly preferred trajectory. The cycle above licenses the sequence

$$h_1 \to h_2 \to h_3 \to h_1,$$

with the agent paying ε at each step. At the end it is back where it started, but poorer by 3ε. Repeating the cycle pumps away arbitrarily much money (Gustafsson, 2010).

There is also an interesting geometric view on transitivity. This paragraph is not needed for the rest of the note, so readers meeting these ideas for the first time can safely treat it as an optional aside. Fix a finite menu $V = \{h_1, \dots, h_m\} \subset \mathcal{H}^*$ and draw the complete graph whose vertices are the trajectories in V. Encode pairwise comparisons by an antisymmetric edge flow X:

$$X_{ij} = -X_{ji},$$

where $X_{ij} > 0$ means that h_i is preferred to h_j, while $X_{ij} < 0$ means the reverse. If preferences come from a utility function u, then each edge weight is just a utility difference,

$$X_{ij} = u(h_i) - u(h_j),$$

so X is a discrete gradient field. In the language of the discrete Helmholtz–Hodge decomposition, any edge flow can be split into a gradient part, which comes from a potential, and a cyclic part, which records genuine loops. On the complete comparison graph there is no extra harmonic remainder, so inconsistency is entirely captured by the cyclic component (Jiang et al., 2011). A three-cycle corresponds exactly to a nonzero discrete curl:

$$X_{ij} + X_{jk} + X_{ki} \neq 0.$$

Thus transitive preferences are precisely the *potential* part of the decomposition, with the potential given by the utility, while intransitive cycles show up as the rotational or cyclic part. If this language feels abstract, the key takeaway is simple: utility means all local comparisons come from one global ranking, whereas cycles are the leftover pattern that cannot be explained by any single scalar potential.

5 Lotteries and the von Neumann–Morgenstern axioms

Up to this point, we have just considered preferences over a countable set $\mathcal{H}^*$ without considering stochasticity. To introduce uncertainty, we enlarge the domain from trajectories to lotteries over trajectories.

Let $\Delta(\mathcal{H}^*)$ denote the set of countably supported probability distributions over $\mathcal{H}^*$. An element $\mu \in \Delta(\mathcal{H}^*)$ can be expressed as

$$\mu = \sum_i p_i \delta_{h_i}$$

which should be read as a lottery that yields trajectory h_i with probability p_i. The trajectory h is identified with the degenerate lottery δ_h. In this way, preferences over trajectories can be viewed as a special case of preferences over lotteries. Furthermore, for $\mu, \nu \in \Delta(\mathcal{H}^*)$ and $\lambda \in [0,1]$, write

$$\lambda\mu + (1-\lambda)\nu$$

for the compound lottery that first flips a coin with bias λ, then samples from μ or ν accordingly.

The von Neumann–Morgenstern (vNM) framework studies weak preference relations $\succcurlyeq$ over $\Delta(\mathcal{H}^*)$ and asks when such preferences admit an expected-utility representation (von Neumann and Morgenstern, 1944). The framework considers four axioms on preferences over lotteries.

Axiom 1 (Completeness)**.** *For all $\mu, \nu \in \Delta(\mathcal{H}^*)$, either $\mu \succcurlyeq \nu$ or $\nu \succcurlyeq \mu$ (or both).*

Axiom 2 (Transitivity)**.** *For all $\mu, \nu, \xi \in \Delta(\mathcal{H}^*)$, if $\mu \succcurlyeq \nu$ and $\nu \succcurlyeq \xi$, then $\mu \succcurlyeq \xi$.*

Axiom 3 (Continuity)**.** *For all $\mu, \nu, \xi \in \Delta(\mathcal{H}^*)$ with $\mu \succcurlyeq \nu \succcurlyeq \xi$, there exists $\lambda \in [0,1]$ such that*

$$\nu \sim \lambda\mu + (1-\lambda)\xi.$$

Axiom 4 (Independence)**.** *For all $\mu, \nu, \xi \in \Delta(\mathcal{H}^*)$ and $\lambda \in (0,1)$,*

$$\mu \succcurlyeq \nu \quad \Longleftrightarrow \quad \lambda\mu + (1-\lambda)\xi \succcurlyeq \lambda\nu + (1-\lambda)\xi.$$

We already know completeness and transitivity from the deterministic setting; the new players are continuity and independence. Continuity says intermediate prospects admit a break-even mixture between better and worse ones. Independence says that if μ is preferred to ν, then mixing both with the same background lottery ξ should not reverse that preference.

The natural question is therefore when a preference over lotteries can be represented by the expectation of a utility function on trajectories:

$$U(\mu) = \sum_{h \in \mathcal{H}^*} \mu(h) u(h).$$

This formulation says that the value of a lottery is the probability-weighted average of the utilities of its possible trajectories. In its simplest finite-outcome form, this question is answered by the celebrated von Neumann–Morgenstern theorem.

Theorem 5.1 (von Neumann–Morgenstern)**.** *Let $X \subseteq \mathcal{H}^*$ be finite. A weak preference relation $\succcurlyeq$ on $\Delta(X)$ satisfies completeness, transitivity, continuity, and independence if and only if there exists a function $u\colon X \to \mathbb{R}$ such that for all lotteries $\mu, \nu \in \Delta(X)$,*

$$\mu \succcurlyeq \nu \quad \Longleftrightarrow \quad \sum_{h \in X} \mu(h) u(h) \geq \sum_{h \in X} \nu(h) u(h).$$

Moreover, u is unique up to positive affine transformations:

$$u'(h) = a + b u(h), \qquad b > 0.$$

In contrast to our previous result, vNM utility is *cardinal* up to positive affine transformations, not merely ordinal. A general monotone transformation would destroy the expectation formula. The independence axiom forces exactly the amount of structure needed for linear averaging.

Two notions of utility.

In economics it is standard to distinguish two different notions of utility (Kreps, 1988; Mas-Colell et al., 1995).

The first is *ordinal utility*, which represents preferences over certain outcomes; in the present note, those outcomes are deterministic trajectories. This is the object obtained in Proposition 4.3: if preferences over deterministic trajectories are complete and transitive, then there exists a function u_{ord} such that

$$h \succcurlyeq h' \quad \Longleftrightarrow \quad u_{\text{ord}}(h) \geq u_{\text{ord}}(h').$$

Only the ranking matters. Any strictly increasing transformation of u_{ord} represents the same preferences (Fishburn, 1970).

The second is *von Neumann–Morgenstern utility*. This is the function u_{vNM} that appears inside the expectation operator when preferences over lotteries satisfy continuity and independence:

$$\mu \succcurlyeq \nu \quad \Longleftrightarrow \quad \sum_h \mu(h) u_{\text{vNM}}(h) \geq \sum_h \nu(h) u_{\text{vNM}}(h).$$

Unlike ordinal utility, u_{vNM} is unique only up to positive affine transformations. Its numerical differences therefore carry behavioral content: they determine which mixtures an agent is willing to accept, and so encode the agent's attitudes toward risk and gambling structure (von Neumann and Morgenstern, 1944; Mas-Colell et al., 1995).

What some expository discussions call *preference utility* ('F1') and *vNM utility* ('F2') is exactly this distinction. 'F1' corresponds to the ordinal representation of certain outcomes; 'F2' corresponds to the cardinal utility recovered from preferences over lotteries. The two agree on the ranking of degenerate outcomes, but 'F2' contains strictly more information than 'F1'.

If continuity or independence fails

It is useful to separate the roles of the four vNM axioms:

- As before, completeness and transitivity give us an ordinal utility on deterministic trajectories.
- Continuity and independence turn that ordinal picture into an *expected* utility theory over uncertain prospects.

If completeness or transitivity fail, as discussed in the previous section, we lose hope of describing the preference by a single scalar quantity. If continuity or independence fails, the preference will generally no longer admit the linear expectation form.

Failure of continuity. To build intuition, let $x \succ y \succ z$ and define

$$\mu_\lambda = \lambda x + (1 - \lambda) z.$$

Then

$$\mu_0 = z, \qquad \mu_1 = x,$$

so as λ increases from 0 to 1, the lottery moves from the worse prospect z to the better prospect x. Continuity says, informally, that preferences vary smoothly along this path. In particular, it says that the intermediate prospect y can be matched by some mixture of the better and worse prospects:

$$y \sim \lambda x + (1-\lambda)z$$

for some $\lambda \in (0,1)$. Intuitively, the axiom requires no abrupt jumps in preference as the mixture probabilities vary.

When continuity fails, some distinctions can become *lexicographic*. For example, we could have a case where $y \succ \mu_0 = z$, but $\mu_\lambda \succ y$ for all $\lambda > 0$. This means that x has absolute priority over y: as soon as x appears with any nonzero probability, the lottery becomes strictly better than y. One way to interpret this is that some considerations are given lexical priority over others. For instance, avoiding catastrophe might outrank ordinary gains so completely that no finite improvement in ordinary reward compensates for even an arbitrarily small increase in catastrophic risk. Such preferences need not be contradictory; they are simply too sharp to be captured by a single real-valued utility whose expectations are taken in the usual way.

What breaks in this case is not all representations, but the *real-valued* vNM representation. If continuity is dropped, one is naturally led to lexicographic or non-Archimedean representations: for example, ordered pairs or vectors of utilities compared lexicographically, or utilities with infinitesimal scales (Fishburn, 1971). In those models, the first coordinate records the highest-priority consideration, and lower coordinates matter only when higher ones tie.

Failure of independence. Independence says that common lottery components should cancel. If $\mu \succcurlyeq \nu$, then mixing both with the same background lottery ξ at the same rate should preserve the ranking:

$$\lambda\mu + (1-\lambda)\xi \succcurlyeq \lambda\nu + (1-\lambda)\xi.$$

So the relative ranking of μ and ν should depend only on how they differ, not on the common part they share.

When independence fails, the value of a prospect depends on its *context*. The same local substitution can be attractive in one background and unattractive in another. This is exactly what the Allais pattern shows (Allais, 1953). Let $h_5 \succ h_1 \succ h_0$ denote trajectories yielding 5, 1, and 0 million, and consider

$$A = 1 \cdot h_1, \qquad B = 0.89\, h_1 + 0.10\, h_5 + 0.01\, h_0,$$

$$C = 0.11\, h_1 + 0.89\, h_0, \qquad D = 0.10\, h_5 + 0.90\, h_0.$$

Many people prefer A to B, but prefer D to C. Under independence this is impossible, because A versus B and C versus D differ only by a common consequence. The reversal shows that certainty is treated as psychologically special: replacing a sure outcome by a tiny risk of getting nothing matters more than expected utility allows.

This is the general lesson of independence failure. Probabilities are no longer aggregated linearly against a fixed utility function on trajectories. Common branches cannot be canceled, and the whole shape of the distribution starts to matter. Decision makers may overweight certainty, distort small probabilities, care about disappointment or regret, or evaluate gains and losses relative to a reference point rather than in absolute terms. This is the route taken by prospect theory and related non-expected-utility models (Kahneman and Tversky, 1979; Machina, 1982).

From the perspective of sequential choice, independence also matters because it allows one to replace a sublottery by an equivalent one without changing the value of the larger plan. If independence fails, the value of a branch may depend on the branches surrounding it, so local and global evaluations need not line up. Hammond's consequentialist argument shows that, together with dynamic consistency and suitable sequential assumptions, one is pushed back toward independence (Hammond, 1988). But that only shows one route to coherent planning. One may instead keep a richer, non-linear evaluation of lotteries and give up the idea that common consequences are always behaviorally irrelevant.

So the two failures have different meanings. Failure of continuity says that some priorities are infinitely sharp, leading naturally to lexicographic or infinitesimal utility scales. Failure of independence says that uncertainty is evaluated holistically rather than by linear averaging, leading to models in which background risk, certainty, or reference dependence affect choice.

6 A fifth axiom: reward and discount

The vNM theorem gives an expected-utility representation over lotteries of whole trajectories, but it does not yet tell us that utility can be generated *locally* from stepwise rewards.

To investigate the possibility of localizing utility over time, let us denote by $T \coloneqq \mathcal{O} \times \mathcal{A}$ the set of one-step transitions. For $t \in T$ and $h \in \mathcal{H}^*$, write $t \cdot h$ for the trajectory obtained by prepending t to h, and extend this operation to lotteries by

$$t \cdot \mu \coloneqq \sum_i p_i \delta_{t \cdot h_i} \qquad \text{when } \mu = \sum_i p_i \delta_{h_i}.$$

Now, using a vNM utility $u(h)$ one can always define a continuation-dependent increment as

$$r(t; h) \coloneqq u(t \cdot h) - u(h).$$

However, this increment may depend on the whole continuation h. What is missing is a temporal condition ensuring that the effect of prepending a transition depends only on that transition itself. Following Bowling et al. (2023), this can be explored with the following additional axiom.

Axiom 5 (Temporal γ-indifference)**.** *There exists a function $\gamma \colon T \to [0, 1]$ such that for all $\mu, \nu \in \Delta(\mathcal{H}^*)$ and all $t \in T$,*

$$\frac{1}{1+\gamma(t)}(t \cdot \mu) + \frac{\gamma(t)}{1+\gamma(t)}\nu \sim \frac{1}{1+\gamma(t)}(t \cdot \nu) + \frac{\gamma(t)}{1+\gamma(t)}\mu.$$

Under expected utility, the indifference above is equivalent to

$$u(t \cdot \mu) + \gamma(t)u(\nu) = u(t \cdot \nu) + \gamma(t)u(\mu),$$

which can be rearranged as

$$u(t \cdot \mu) - u(t \cdot \nu) = \gamma(t)\big(u(\mu) - u(\nu)\big).$$

So the effect of prepending the same transition t is to rescale the utility difference between two continuations by a factor $\gamma(t)$. In that sense, $\gamma(t)$ measures how much the future still matters after the step t has occurred. When $\gamma(t) = 1$, future differences are preserved without discount at that step. When $0 < \gamma(t) < 1$, the future still matters but is damped. When $\gamma(t) = 0$, once t has occurred the continuation no longer affects the comparison. Bowling's axiom is closely related to the transition-dependent discounting discussed by White (2017) and Pitis (2019).

Bowling et al. show that adding this axiom to the four vNM axioms is necessary and sufficient for a discounted-reward representation of preferences.

Theorem 6.1 (Markov reward representation, after Bowling et al.)**.** *A weak preference relation $\succcurlyeq$ on $\Delta(\mathcal{H}^*)$ satisfies completeness, transitivity, continuity, independence, and temporal γ-indifference if and only if there exist functions $u\colon \mathcal{H}^* \to \mathbb{R}$, $r\colon T \to \mathbb{R}$, and $\gamma\colon T \to [0,1]$ such that $u(\varepsilon) = 0$,*

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

and, for all lotteries $\mu, \nu \in \Delta(\mathcal{H}^*)$,

$$\mu \succcurlyeq \nu \quad \Longleftrightarrow \quad \sum_{h \in \mathcal{H}^*} \mu(h)u(h) \geq \sum_{h \in \mathcal{H}^*} \nu(h)u(h).$$

Moreover, r is unique up to positive scale, and γ is the same function appearing in the fifth axiom (Bowling et al., 2023).

This theorem sharpens the vNM result in exactly the way reinforcement learning needs. Utility is no longer an arbitrary scalar attached to a complete trajectory; it is generated recursively from a local reward and a local weight on the future. Unrolling the recursion for a trajectory $h = (t_1, t_2, \ldots, t_n)$ gives

$$u(h) = r(t_1) + \gamma(t_1)r(t_2) + \gamma(t_1)\gamma(t_2)r(t_3) + \cdots + \left(\prod_{i=1}^{n-1} \gamma(t_i)\right) r(t_n).$$

Thus the fifth axiom is what allows utility over whole trajectories to be decomposed into local reward and discounting. Two familiar special cases are worth highlighting.

- If $\gamma(t) \equiv \gamma$ is constant, then
$$u(h) = \sum_{k=1}^{n} \gamma^{k-1} r(t_k),$$
which is the standard discounted-return objective.
- If $\gamma(t) \equiv 1$, then
$$u(h) = \sum_{k=1}^{n} r(t_k),$$
so utility is simply the additive cumulative reward.

In summary, the step from vNM utility over whole trajectories to RL-style reward requires one more axiom. That axiom simultaneously identifies both the local reward signal and the form of discounting.

Remark 6.2 (MDPs, POMDPs, and locality)**.** *The locality result of this section is especially natural in fully observed MDPs, where one often expects reward to depend only on the current state transition. In the present note, however, the primitive histories are sequences of observations and actions, and the corresponding local reward has the form $r\colon \mathcal{O} \times \mathcal{A} \to \mathbb{R}$. In partially observed settings this can be restrictive: a goal may be Markov in the hidden state, in the agent's belief state, or in some augmented memory state, without being reducible to a function of the current observation-action pair alone. In that sense, the fifth axiom should be read as characterizing when preferences admit a reward that is local in the chosen representation of experience. If the raw observation stream is too coarse, one may need to enrich the state description before a Markov reward representation becomes available (Bowling et al., 2023).*

Utility is not reward.
It is helpful to keep the following four different objects apart.

- *Preference* is the primitive relation $\succcurlyeq$. It says only which trajectories or lotteries are weakly preferred to which others.
- *Preference utility* ('F1') is an ordinal representation of preferences over deterministic trajectories. Under completeness and transitivity, it is any function u_{ord} such that

$$h \succcurlyeq h' \quad \Longleftrightarrow \quad u_{\text{ord}}(h) \geq u_{\text{ord}}(h').$$

  It is unique only up to strictly increasing transformations (Fishburn, 1970).
- *vNM utility* ('F2') is the stronger, cardinal utility that appears when preferences over lotteries satisfy the vNM axioms. It is the function u_{vNM} for which

$$\mu \succcurlyeq \nu \quad \Longleftrightarrow \quad \sum_h \mu(h) u_{\text{vNM}}(h) \geq \sum_h \nu(h) u_{\text{vNM}}(h).$$

  It is unique only up to positive affine transformations (von Neumann and Morgenstern, 1944). Thus 'F2' refines 'F1': it agrees with it on the ranking of certain trajectories, but adds the extra structure needed to compare lotteries.
- *Reward* is not either of these utilities. A reward function $r(t)$ is a *local* representation introduced only after adding temporal structure. In the present framework, reward appears when the fifth axiom allows utility to be written recursively as

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

  So reward is a way of *decomposing* utility over complete trajectories into stepwise contributions. It is therefore downstream of preference, and even downstream of vNM utility. This is why different reward functions can encode the same utility or the same preference ordering, a point we return to in the next section (Bowling et al., 2023).

7 Reward equivalence and shaping

The preceding theorem shows how to pass from utility over lotteries of trajectories to a local reward representation. But it also raises an important question: how unique is that reward? The answer has two parts. If one fixes a particular utility function u and a particular discount function γ, then the reward is determined. But if one changes the numerical representative of the same preference relation, or allows shaping transformations, then many different rewards can encode the same underlying ordering.

First observe that once u and γ are fixed, the reward is fixed as well. Indeed, if

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

for all $t \in T$ and $h \in \mathcal{H}^*$, then necessarily

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

and the right-hand side must be independent of the continuation h. So the non-uniqueness of reward does not come from ambiguity inside a fixed recursive representation. It comes from the fact that utility itself is not unique as a numerical object.

Proposition 7.1 (Multiple rewards can encode the same preference relation)**.** *Suppose $u\colon \mathcal{H}^* \to \mathbb{R}$, $r\colon T \to \mathbb{R}$, and $\gamma\colon T \to [0,1]$ satisfy*

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

for all $t \in T$ and $h \in \mathcal{H}^*$, and that preferences over lotteries are represented by the expected value of u. For any $a \in \mathbb{R}$ and $b > 0$, define

$$u'(h) = a + bu(h) \qquad \text{and} \qquad r'(t) = br(t) + (1 - \gamma(t))a.$$

Then $u'(t \cdot h) = r'(t) + \gamma(t)u'(h)$ for all $t \in T$ and $h \in \mathcal{H}^*$, and the expected value of u' represents the same preference relation as the expected value of u.

The proposition shows that reward non-uniqueness already appears at the level of affine reparameterizations of utility: changing the numerical representative of the same vNM preference relation generally changes the associated reward function as well. This point is especially simple under the normalization $u(\varepsilon) = 0$ used in the previous section. Then the additive degree of freedom disappears, and one is left only with positive rescalings:

$$u'(h) = bu(h), \qquad r'(t) = br(t).$$

So in the normalized setting reward is unique only up to choice of units.

There is also a second, less trivial kind of non-uniqueness, in which one changes rewards while leaving the induced trajectory ordering unchanged, first studied by Ng et al. (1999).

Proposition 7.2 (Potential-based shaping changes utility only by a boundary term)**.** *Consider a state-based setting with transitions $t = (s, a, s')$ and a potential function Φ on states.*

(a) If $\gamma \equiv 1$ and $r_\Phi(s,a,s') \coloneqq r(s,a,s') + \Phi(s') - \Phi(s)$, then

$$\sum_{k=1}^{n} r_\Phi(t_k) = \sum_{k=1}^{n} r(t_k) + \Phi(s_n) - \Phi(s_0)$$

for any trajectory $\tau = (t_1, \ldots, t_n)$ from s_0 to s_n.

(b) If $\gamma \in (0,1)$ is constant and $r_\Phi(s,a,s') \coloneqq r(s,a,s') + \gamma\Phi(s') - \Phi(s)$, then

$$\sum_{k=1}^{n} \gamma^{k-1} r_\Phi(t_k) = \sum_{k=1}^{n} \gamma^{k-1} r(t_k) - \Phi(s_0) + \gamma^n \Phi(s_n).$$

The previous result shows that shaping r into r_Φ changes utility only by a boundary term. In particular, if all compared trajectories share the same start state and terminal potential, then the induced preference ordering is unchanged; if the boundary term vanishes on all admissible trajectories, then even the numerical utility is unchanged.

The general lesson is that reward is best understood as a *coordinate system* for utility, not as a uniquely given primitive. Some transformations, such as positive scaling, merely change the numerical units. Others, such as potential-based shaping, redistribute value along the trajectory while leaving the overall preference over complete trajectories unchanged. This is why reward design is often underdetermined: what matters behaviorally is not a raw reward function in isolation, but the utility and preference structure it induces.

8 Conclusion

The main lesson of this note is that, in sequential settings, the primitive object is not a one-step reward but a preference over complete trajectories. From that starting point one can distinguish several layers of structure. Completeness and transitivity yield an ordinal utility over deterministic trajectories. Once lotteries are introduced, continuity and independence refine that ordinal picture into a von Neumann–Morgenstern utility, which agrees with the ranking of certain trajectories but carries strictly more information because it calibrates trade-offs between risky prospects.

This separation of layers clarifies both what expected utility achieves and what it leaves out. It explains why "utility" in economics can refer either to an ordinal representation of certain preferences or to a cardinal object suitable for evaluating lotteries. It also shows what fails when particular axioms are dropped: without completeness or transitivity, scalar representation itself can fail; without continuity or independence, one may still have meaningful preferences, but no longer the linear expectation form of vNM. In that sense, expected utility is not the whole of rationality, but a particular and mathematically powerful strengthening of it.

The final step is to see that even vNM utility is not yet reward. To obtain a local reward-and-discount representation one needs additional temporal structure, captured here by Bowling et al.'s fifth axiom. Under that axiom, utility over whole trajectories admits a recursive decomposition into local reward and discounting. But even then reward is not unique: affine changes of utility induce corresponding changes of reward, and potential-based shaping can redistribute value along a trajectory while preserving the underlying ordering. Moreover, the locality of reward depends on the chosen representation of experience, which is natural in fully observed MDPs but can be restrictive in partially observed settings unless the state is suitably enriched. The reward hypothesis, understood in this way, is therefore best read not as the claim that goals are primitively rewards, but as the claim that sufficiently structured preferences over trajectories can be represented by rewards.

References

Maurice Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'ecole americaine. *Econometrica*, 21(4):503–546, 1953.

Robert J Aumann. Utility theory without the completeness axiom. *Econometrica*, 30(3):445–462, 1962.

Michael Bowling, John D. Martin, David Abel, and Will Dabney. Settling the reward hypothesis. *arXiv preprint arXiv:2212.10420*, 2023. URL https://arxiv.org/abs/2212.10420.

Gerard Debreu. Representation of a preference ordering by a numerical function. In Robert M. Thrall, Clyde H. Coombs, and Ralph L. Davis, editors, *Decision Processes*, pages 159–165. Wiley, New York, 1954.

Peter C. Fishburn. *Utility Theory for Decision Making*. Wiley, New York, 1970.

Peter C. Fishburn. A study of lexicographic expected utility. *Management Science*, 17(11):672–678, 1971. doi: 10.1287/mnsc.17.11.672.

Johan E. Gustafsson. A money-pump for acyclic intransitive preferences. *Dialectica*, 64(2):251–257, 2010. doi: 10.1111/j.1746-8361.2010.01230.x.

Peter J. Hammond. Consequentialist foundations for expected utility. *Theory and Decision*, 25(1): 25–78, 1988. doi: 10.1007/BF00129168.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011. doi: 10.1007/s10107-010-0419-x.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

David M. Kreps. *Notes on the Theory of Choice*. Westview Press, Boulder, CO, 1988.

Mark J. Machina. "expected utility" analysis without the independence axiom. *Econometrica*, 50(2): 277–323, 1982.

Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, New York, 1995.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287, 1999.

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

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

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Martha White. Unifying task specification in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Exercises

The following exercises are meant to help students reconstruct the main results of the note for themselves. The last two are guided proof exercises for the two central representation theorems.

Exercise 1 (Ordinal versus vNM utility)**.** *Let $x, y, z \in \mathcal{H}^*$ satisfy $x \succ y \succ z$.*

(a) Give two different ordinal utility functions $u_1, u_2 \colon \mathcal{H}^* \to \mathbb{R}$ that represent the same ranking of x, y, z.

(b) Explain why these two functions are equally good as ordinal representations.

(c) Suppose in addition that

$$y \sim \tfrac{1}{2}x + \tfrac{1}{2}z.$$

Show that not every strictly increasing transformation of a vNM utility can preserve this indifference.

Exercise 2 (Failures of the vNM axioms)**.** *Here we will study the consequences of different axioms.*

(a) Construct a simple incomplete preference relation on three trajectories. Why can it not be represented by a single real-valued ordinal utility?

(b) Construct a three-cycle and explain how it gives rise to a money pump.

(c) Give an example of a lexicographic preference over three outcomes and show that it violates continuity.

(d) Write down the Allais pattern from section 5 and explain which axiom it violates.

Exercise 3 (Reward shaping)**.** *Assume additive reward with $\gamma \equiv 1$ and define*

$$r_\Phi(s, a, s') = r(s, a, s') + \Phi(s') - \Phi(s).$$

(a) Show that along any finite trajectory $(t_1, \ldots, t_n)$ the total shaped reward differs from the original total reward by the boundary term $\Phi(s_n) - \Phi(s_0)$.

(b) Under what condition on the admissible trajectories does this shaping leave utility exactly unchanged?

(c) Under what weaker condition does it leave only the induced preference ordering unchanged?

Exercise 4 (Guided proof of the von Neumann–Morgenstern theorem)**.** *Assume $X \subseteq \mathcal{H}^*$ is finite and that $\succcurlyeq$ on $\Delta(X)$ satisfies completeness, transitivity, continuity, and independence.*

(a) Show that there exist trajectories $b, w \in X$ such that $b \succcurlyeq h \succcurlyeq w$ for every $h \in X$.

(b) Using continuity, show that for every $h \in X$ there exists $u(h) \in [0, 1]$ such that

$$h \sim u(h)b + (1 - u(h))w.$$

(c) Use transitivity to show that if

$$h \sim u(h)b + (1 - u(h))w \quad \text{and} \quad h' \sim u(h')b + (1 - u(h'))w,$$

then

$$h \succcurlyeq h' \quad \Longleftrightarrow \quad u(h) \geq u(h').$$

(d) Let

$$\mu = \sum_i p_i h_i.$$

Use independence repeatedly to replace each h_i by the equivalent lottery $u(h_i)b + (1 - u(h_i))w$ and show that

$$\mu \sim \sum_i p_i \big(u(h_i)b + (1 - u(h_i))w\big).$$

(e) Reduce the compound lottery in part (d) to a simple lottery and prove that

$$\mu \sim U(\mu)b + (1 - U(\mu))w, \qquad U(\mu) \coloneqq \sum_i p_i u(h_i).$$

(f) Show that

$$\mu \succcurlyeq \nu \quad \iff \quad U(\mu) \geq U(\nu).$$

This gives the expected-utility representation.

(g) Prove affine uniqueness: if u' also represents the same preference relation in expected-utility form, then $u' = a + bu$ for some $a \in \mathbb{R}$ and $b > 0$.

(h) Prove the converse direction: if preferences admit an expected-utility representation, then they satisfy completeness, transitivity, continuity, and independence.

Exercise 5 (Guided proof of the Markov reward representation theorem)**.** *Assume the hypotheses of the previous exercise together with temporal γ-indifference, and let u be a vNM utility representation over trajectories.*

(a) Apply expected utility to temporal γ-indifference and show that for all lotteries $\mu, \nu \in \Delta(\mathcal{H}^*)$ and all transitions $t \in T$,

$$u(t \cdot \mu) + \gamma(t)u(\nu) = u(t \cdot \nu) + \gamma(t)u(\mu).$$

(b) Rearrange part (a) to prove that

$$u(t \cdot \mu) - \gamma(t)u(\mu)$$

is independent of the choice of μ.

(c) Define

$$r(t) \coloneqq u(t \cdot \mu) - \gamma(t)u(\mu)$$

for any μ. Deduce that for every deterministic trajectory h,

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

(d) Show that this recursion extends to lotteries by linearity:

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

(e) Prove the converse direction: if u, r, and γ satisfy

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

for all t and h, and if u extends linearly to lotteries, then temporal γ-indifference holds.

(f) Unroll the recursion along a finite trajectory $h = (t_1, \ldots, t_n)$ and derive

$$u(h) = r(t_1) + \gamma(t_1)r(t_2) + \gamma(t_1)\gamma(t_2)r(t_3) + \cdots + \left(\prod_{i=1}^{n-1} \gamma(t_i)\right) r(t_n).$$

(g) Show that when $\gamma(t) \equiv \gamma$ is constant, the previous formula reduces to standard discounted return, and when $\gamma(t) \equiv 1$ it reduces to additive cumulative reward.

(h) Show that if $u' = a + bu$, then the corresponding reward must be

$$r'(t) = br(t) + (1 - \gamma(t))a.$$

Interpret this as a source of reward non-uniqueness.

Open-ended questions

Exercise 6. *Do the results have any prescriptive or descriptive implications? What kinds of agents, with what kinds of preferences should we design? By default, what kind of agents should we expect to obtain via typical training processes?*

Exercise 7. *Let's assume a superintelligence has preferences over trajectories. Are there weaker versions of the axioms that make it more plausible that these preferences are "safe for us" compared to preferences that lead to utilities or even rewards?*

Exercise 8. *More generally, are there interesting examples of preferences over trajectories that the students could analyse, that do or do not satisfy some of the axioms?*