From preferences to rewards: A brief introduction

Fernando E. Rosas

Iliad Intensive

September 8, 2026

The Question

- Agents are powerful — but increasing capabilities lead to increasing risks.

The Question

- Agents are powerful — but increasing capabilities lead to increasing risks.
- To understand agentic risk, we need to understand how agents work.

The Question

- Agents are powerful — but increasing capabilities lead to increasing risks.
- To understand agentic risk, we need to understand how agents work.
- Core of agency: *doing things (planning) for reasons (beliefs)*.

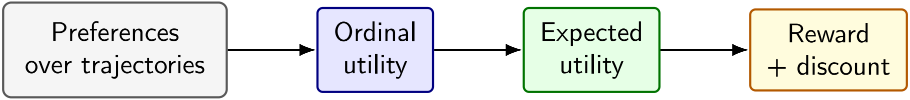
The Question

- Agents are powerful — but increasing capabilities lead to increasing risks.
- To understand agentic risk, we need to understand how agents work.
- Core of agency: *doing things (planning) for reasons (beliefs)*.
- What are goals? Hard question.

The Question

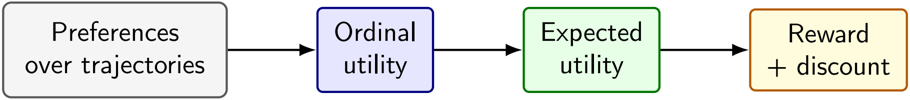
- Agents are powerful — but increasing capabilities lead to increasing risks.
- To understand agentic risk, we need to understand how agents work.
- Core of agency: *doing things (planning) for reasons (beliefs)*.
- What are goals? Hard question.
- **Goal of today**: understand what are preferences.

The Route



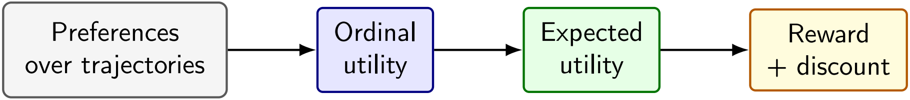
- Each arrow adds structure.

The Route



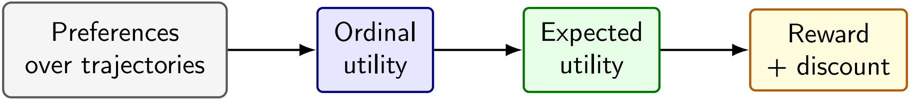
- Each arrow adds structure.
- Each structure buys a more specific representation.

The Route



- Each arrow adds structure.
- Each structure buys a more specific representation.
- *Outcome 1*: Understand the assumptions of each step.

The Route



- Each arrow adds structure.
- Each structure buys a more specific representation.
- *Outcome 1*: Understand the assumptions of each step.
- *Outcome 2*: Recognise the difference between preference, utility, and reward.

Basic setting

- Let $\mathcal{O}$ be a finite observation set, and $\mathcal{A}$ a finite action set.

Basic setting

- Let $\mathcal{O}$ be a finite observation set, and $\mathcal{A}$ a finite action set.
- A one-step interaction is

$$t = (o, a) \in \mathcal{O} \times \mathcal{A}.$$

Basic setting

- Let $\mathcal{O}$ be a finite observation set, and $\mathcal{A}$ a finite action set.
- A one-step interaction is

$$t = (o, a) \in \mathcal{O} \times \mathcal{A}.$$

- A finite trajectory of length n is

$$h = (o_1, a_1, o_2, a_2, \ldots, o_n, a_n) \in (\mathcal{O} \times \mathcal{A})^n.$$

Basic setting

- Let $\mathcal{O}$ be a finite observation set, and $\mathcal{A}$ a finite action set.
- A one-step interaction is

$$t = (o, a) \in \mathcal{O} \times \mathcal{A}.$$

- A finite trajectory of length n is

$$h = (o_1, a_1, o_2, a_2, \ldots, o_n, a_n) \in (\mathcal{O} \times \mathcal{A})^n.$$

- The full finite trajectory space is

$$\mathcal{H}^* = \bigcup_{n=0}^{\infty} (\mathcal{O} \times \mathcal{A})^n.$$

Preferences

- A weak preference relation $\succcurlyeq$ compares complete trajectories.

$$h \succcurlyeq h'$$

means "h is at least as good as h'."

Preferences

- A weak preference relation $\succcurlyeq$ compares complete trajectories.

$$h \succcurlyeq h'$$

  means "h is at least as good as h'."
- From this we define:

$$\begin{aligned} h \sim h' &\iff h \succcurlyeq h' \text{ and } h' \succcurlyeq h, \\ h \succ h' &\iff h \succcurlyeq h' \text{ and not } h' \succcurlyeq h. \end{aligned}$$

Preferences

- A weak preference relation $\succcurlyeq$ compares complete trajectories.

$$h \succcurlyeq h'$$

  means "h is at least as good as h'."
- From this we define:

$$\begin{aligned} h \sim h' &\iff h \succcurlyeq h' \text{ and } h' \succcurlyeq h, \\ h \succ h' &\iff h \succcurlyeq h' \text{ and not } h' \succcurlyeq h. \end{aligned}$$

- At this point, $\succcurlyeq$ need not be complete, transitive, numerical, or 'reward-like'.

Three Kinds of Trouble

By choosing the "wrong type of preference", an agent can run into three kinds of trouble (from less to more harmful):

- **Representability failure:** preference has no convenient formal representation.

Three Kinds of Trouble

By choosing the "wrong type of preference", an agent can run into three kinds of trouble (from less to more harmful):

- **Representability failure:** preference has no convenient formal representation.
- **Self-defeat:** agents take suboptimal choices or plans — there are better alternatives that are not being taken.

Three Kinds of Trouble

By choosing the "wrong type of preference", an agent can run into three kinds of trouble (from less to more harmful):

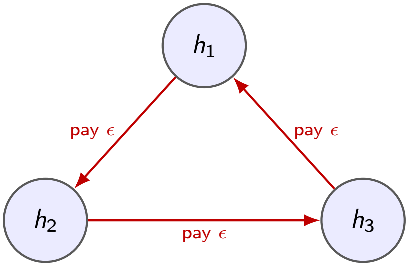
- **Representability failure:** preference has no convenient formal representation.
- **Self-defeat:** agents take suboptimal choices or plans — there are better alternatives that are not being taken.
  - E.g. **dominance**: another policy yields preferred trajectories at least as preferred in every state of the world, and strictly so in at least one.

Three Kinds of Trouble

By choosing the "wrong type of preference", an agent can run into three kinds of trouble (from less to more harmful):

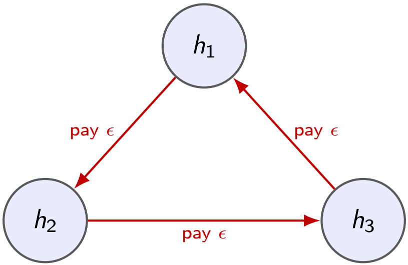
- **Representability failure:** preference has no convenient formal representation.
- **Self-defeat:** agents take suboptimal choices or plans — there are better alternatives that are not being taken.
  - E.g. **dominance**: another policy yields preferred trajectories at least as preferred in every state of the world, and strictly so in at least one.
- **Vulnerability:** individually acceptable trades can combine into guaranteed loss.
  - E.g. **Money pumps**: circular preferences ($h_1 \succ h_2 \succ h_3 \succ h_1$) lead to indefinite loss.

Money Pump Picture



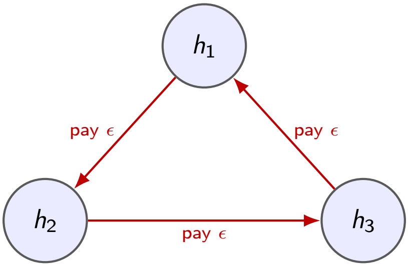
- If the agent pays for every strict improvement, a cycle can be exploited.

Money Pump Picture



- If the agent pays for every strict improvement, a cycle can be exploited.
- It returns to the same trajectory with less money.

Money Pump Picture



- If the agent pays for every strict improvement, a cycle can be exploited.
- It returns to the same trajectory with less money.
- Parallel with thermodynamics: the second law of thermodynamics can be read as stating that *'nature cannot be money-pumped'*.

Three Kinds of Results

There are three types of results regarding preferences:

- **Representation results:** axioms imply that preferences can be written in a particular mathematical form.
  - E.g. some preferences can be represented as if maximising a utility function.

Three Kinds of Results

There are three types of results regarding preferences:

- **Representation results:** axioms imply that preferences can be written in a particular mathematical form.
  - E.g. some preferences can be represented as if maximising a utility function.
- **Coherence results:** violations imply some penalty for the agent.
  - E.g. money pumps or Dutch books.

Three Kinds of Results

There are three types of results regarding preferences:

- **Representation results:** axioms imply that preferences can be written in a particular mathematical form.
  - ▷ E.g. some preferences can be represented as if maximising a utility function.
- **Coherence results:** violations imply some penalty for the agent.
  - ▷ E.g. money pumps or Dutch books.
- **Selection results:** some optimization pressure tends to remove agents with a given failure mode.
  - ▷ E.g. market competition or natural evolution.

Three Kinds of Results

There are three types of results regarding preferences:

- **Representation results:** axioms imply that preferences can be written in a particular mathematical form.
  - ▷ E.g. some preferences can be represented as if maximising a utility function.
- **Coherence results:** violations imply some penalty for the agent.
  - ▷ E.g. money pumps or Dutch books.
- **Selection results:** some optimization pressure tends to remove agents with a given failure mode.
  - ▷ E.g. market competition or natural evolution.

These results are related (one can lead to the other) but not identical: form, vulnerability, and survival are different claims.

Axiom 1: Completeness

Completeness says every pair can be compared:

$$\forall h, h' \in \mathcal{H}^*, \qquad h \succcurlyeq h' \quad \text{or} \quad h' \succcurlyeq h.$$

Axiom 1: Completeness

Completeness says every pair can be compared:

$$\forall h, h' \in \mathcal{H}^*, \qquad h \succcurlyeq h' \quad \text{or} \quad h' \succcurlyeq h.$$

- Incomparability is not the same as indifference.

Axiom 1: Completeness

Completeness says every pair can be compared:

$$\forall h, h' \in \mathcal{H}^*, \qquad h \succcurlyeq h' \quad \text{or} \quad h' \succcurlyeq h.$$

- Incomparability is not the same as indifference.
- If it fails, some trajectories are genuinely incomparable.

Axiom 2: Transitivity

Transitivity says comparisons fit together:

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

Axiom 2: Transitivity

Transitivity says comparisons fit together:

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

- If it fails, local pairwise judgments need not form a global ranking.

Axiom 2: Transitivity

Transitivity says comparisons fit together:

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

- If it fails, local pairwise judgments need not form a global ranking.
- The classic failure is a cycle: money pumps.

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

- Similar to Helmholtz-Hodge decomposition: transitivity is equivalent to requiring no rotational component.

Axioms 1 + 2 $\rightarrow$ Ordinal Utility

- Completeness + transitivity make $\succcurlyeq$ a weak order (total order with equivalent objects).

Axioms 1 + 2 $\rightarrow$ Ordinal Utility

- Completeness + transitivity make $\succcurlyeq$ a weak order (total order with equivalent objects).
- On a countable space $\mathcal{H}^*$, weak orders admit an ordinal utility: there exists a function $u : \mathcal{H}^* \to \mathbb{R}$ such that

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

Axioms 1 + 2 $\rightarrow$ Ordinal Utility

- Completeness + transitivity make $\succcurlyeq$ a weak order (total order with equivalent objects).
- On a countable space $\mathcal{H}^*$, weak orders admit an ordinal utility: there exists a function $u : \mathcal{H}^* \rightarrow \mathbb{R}$ such that

$$h \succcurlyeq h' \iff u(h) \geq u(h').$$

- Only the ranking matters:

$$u, \quad 2u + 17, \quad \exp(u)$$

all represent the same deterministic preferences.

Beyond Ordinal Utility

Ordinal utility ranks outcomes, but gives a structure to mixtures of them.

- Example:

$$x \succ y \succ z.$$

  How should we rank y with respect to $\frac{1}{2}x + \frac{1}{2}z$? Should they be equivalent?

Beyond Ordinal Utility

Ordinal utility ranks outcomes, but gives a structure to mixtures of them.

- Example:

$$x \succ y \succ z.$$

  How should we rank y with respect to $\frac{1}{2}x + \frac{1}{2}z$? Should they be equivalent?

Ordinal rank doesn't say how to rank mixtures, which can arise naturally out of uncertainty.

Lotteries Over Trajectories

- Let $\Delta(X)$ be the set of probability distributions over $X \subseteq \mathcal{H}^*$.

Lotteries Over Trajectories

- Let $\Delta(X)$ be the set of probability distributions over $X \subseteq \mathcal{H}^*$.
- A lottery is written

$$\mu = \sum_i p_i \delta_{x_i}.$$

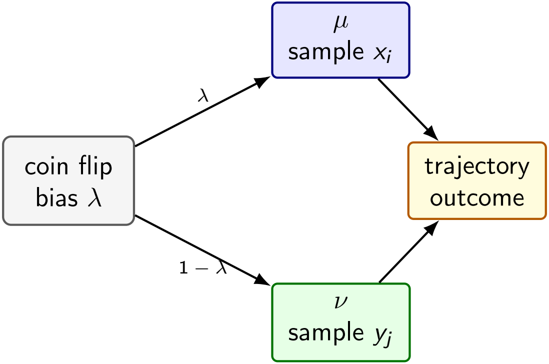
Lotteries Over Trajectories

- Let $\Delta(X)$ be the set of probability distributions over $X \subseteq \mathcal{H}^*$.
- A lottery is written
$$\mu = \sum_i p_i \delta_{x_i}.$$
- A sure trajectory x is the degenerate lottery δ_x.

Lotteries Over Trajectories

- Let $\Delta(X)$ be the set of probability distributions over $X \subseteq \mathcal{H}^*$.
- A lottery is written
$$\mu = \sum_i p_i \delta_{x_i}.$$
- A sure trajectory x is the degenerate lottery δ_x.
- We can mix lotteries:
$$\lambda\mu + (1-\lambda)\nu.$$

Lottery Picture



- vNM asks when preferences over these objects are expectations.

vNM representation: The big picture

1. **Axiom 1: Completeness:** any two lotteries can be compared.

vNM representation: The big picture

1. **Axiom 1: Completeness:** any two lotteries can be compared.
2. **Axiom 2: Transitivity:** comparisons over lotteries fit together.

vNM representation: The big picture

1. **Axiom 1: Completeness:** any two lotteries can be compared.
2. **Axiom 2: Transitivity:** comparisons over lotteries fit together.
3. **Axiom 3: Continuity:** intermediate prospects can be matched by a mixture of better and worse prospects.

vNM representation: The big picture

1. **Axiom 1: Completeness:** any two lotteries can be compared.
2. **Axiom 2: Transitivity:** comparisons over lotteries fit together.
3. **Axiom 3: Continuity:** intermediate prospects can be matched by a mixture of better and worse prospects.
4. **Axiom 4: Independence:** common lottery components cancel.

vNM representation: The big picture

1. **Axiom 1: Completeness:** any two lotteries can be compared.
2. **Axiom 2: Transitivity:** comparisons over lotteries fit together.
3. **Axiom 3: Continuity:** intermediate prospects can be matched by a mixture of better and worse prospects.
4. **Axiom 4: Independence:** common lottery components cancel.

Result: The first two axioms give ranking structure based on an ordinal utility; the last two axioms make the utility an expected value.

Axiom 3: Continuity

- If $\mu \succcurlyeq \nu \succcurlyeq \xi$, continuity says there always exists some value $\lambda \in [0, 1]$ such that
$$\nu \sim \lambda\mu + (1 - \lambda)\xi.$$

Axiom 3: Continuity

- If $\mu \succcurlyeq \nu \succcurlyeq \xi$, continuity says there always exists some value $\lambda \in [0, 1]$ such that
$$\nu \sim \lambda\mu + (1 - \lambda)\xi.$$
- Intuition: no outcome is infinitely better or worse than the others.

Axiom 3: Continuity

- If $\mu \succcurlyeq \nu \succcurlyeq \xi$, continuity says there always exists some value $\lambda \in [0, 1]$ such that
$$\nu \sim \lambda\mu + (1 - \lambda)\xi.$$

- Intuition: no outcome is infinitely better or worse than the others.

Classic counter-examples are lexicographic preferences, in which even the slimmest chance of a catastrophically bad outcome is worse than a slightly bad outcome with certianty, no matter how low the odds are.

Example: μ is eating chocolate, ν is eating potatoes, and ξ is accepting genocide.

Axiom 4: Independence

- If $\mu \succcurlyeq \nu$, then mixing both with the same background lottery ξ should not reverse the comparison:

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

Axiom 4: Independence

- If $\mu \succcurlyeq \nu$, then mixing both with the same background lottery ξ should not reverse the comparison:

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

- Independence is what gives linear averaging.

Axiom 4: Independence

- If $\mu \succcurlyeq \nu$, then mixing both with the same background lottery ξ should not reverse the comparison:

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

- Independence is what gives linear averaging.

When it fails, common consequences do not psychologically or behaviorally cancel.

Example (Allais paradox): Most people would say

- Getting \$100 for sure *is better than* a coin flip for \$250.
- A 1-in-100 shot at \$100 *is worse than* a 1-in-200 shot at \$250.

These are the same choice, just scaled down.

Theorem 1: The von Neumann–Morgenstern representation

Assumptions: a finite set of options $X \subseteq \mathcal{H}^*$, the four axioms hold.

- **Result 1 (existence):** there exists $u : X \to \mathbb{R}$ such that

$$\mu \succcurlyeq \nu \iff \sum_{x \in X} \mu(x) u(x) \geq \sum_{x \in X} \nu(x) u(x).$$

Theorem 1: The von Neumann–Morgenstern representation

Assumptions: a finite set of options $X \subseteq \mathcal{H}^*$, the four axioms hold.

- **Result 1 (existence):** there exists $u : X \to \mathbb{R}$ such that

$$\mu \succcurlyeq \nu \iff \sum_{x \in X} \mu(x)u(x) \geq \sum_{x \in X} \nu(x)u(x).$$

- **Result 2 (uniqueness):** u is unique up to positive affine transformations:

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

Ordinal vs expected utility

Ordinal utility

- Any set of outcomes
- Preserves ranking
- Indiferent to non-decreasing transformation
- Cannot display ‘meaningful gaps’

vNM utility

- Lotteries (probabilistic mixtures)
- Preserves risky choices
- Only affine transforms are okay
- Incommensurable gaps can exist

Ordinal vs expected utility

Ordinal utility

- Any set of outcomes
- Preserves ranking
- Indiferent to non-decreasing transformation
- Cannot display 'meaningful gaps'

vNM utility

- Lotteries (probabilistic mixtures)
- Preserves risky choices
- Only affine transforms are okay
- Incommensurable gaps can exist

vNM utility further refines ordinal utility (it is not just a relabeling).

Beyond expected utility: Temporal structure

- vNM utility assigns a single scalar to a whole trajectory.

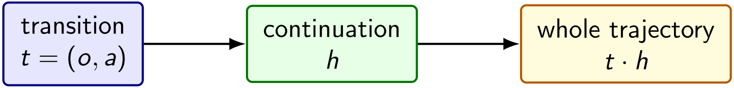
Beyond expected utility: Temporal structure

- vNM utility assigns a single scalar to a whole trajectory.
- But agents make decisions over time, so they need indications of utility that unfold step-by-step.

Beyond expected utility: Temporal structure

- vNM utility assigns a single scalar to a whole trajectory.
- But agents make decisions over time, so they need indications of utility that unfold step-by-step.
- From a global utility $u(h)$, one can always define the increment

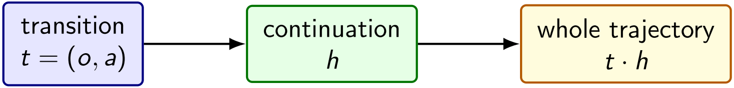
$$r(t;h) = u(t \cdot h) - u(h), \quad \text{satisfying} \quad u(t \cdot h) = r(t;h) + u(h).$$



Beyond expected utility: Temporal structure

- vNM utility assigns a single scalar to a whole trajectory.
- But agents make decisions over time, so they need indications of utility that unfold step-by-step.
- From a global utility $u(h)$, one can always define the increment

$$r(t; h) = u(t \cdot h) - u(h), \quad \text{satisfying} \quad u(t \cdot h) = r(t; h) + u(h).$$



- But the increment $r(t; h)$ may be non-local, and hence cannot be optimised for directly.

Axiom 5: Temporal equanimity

Let $T = \mathcal{O} \times \mathcal{A}$ be the transition set.

- The fifth axiom says there is a function $\gamma : T \to [0, 1]$ such that prepending t rescales continuation differences, so that

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

$\gamma(t)$ measures how much the future still matters after t.

Axiom 5: Temporal equanimity

Let $T = \mathcal{O} \times \mathcal{A}$ be the transition set.

- The fifth axiom says there is a function $\gamma : T \to [0,1]$ such that prepending t rescales continuation differences, so that

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

$\gamma(t)$ measures how much the future still matters after t.

- $\gamma(t) = 1$: future differences matter as much as present differences.

Axiom 5: Temporal equanimity

Let $T = \mathcal{O} \times \mathcal{A}$ be the transition set.

- The fifth axiom says there is a function $\gamma : T \to [0, 1]$ such that prepending t rescales continuation differences, so that

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

$\gamma(t)$ measures how much the future still matters after t.

- $\gamma(t) = 1$: future differences matter as much as present differences.
- $0 < \gamma(t) < 1$: future differences are damped.

Axiom 5: Temporal equanimity

Let $T = \mathcal{O} \times \mathcal{A}$ be the transition set.

- The fifth axiom says there is a function $\gamma : T \to [0,1]$ such that prepending t rescales continuation differences, so that

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

$\gamma(t)$ measures how much the future still matters after t.

- $\gamma(t) = 1$: future differences matter as much as present differences.
- $0 < \gamma(t) < 1$: future differences are damped.
- $\gamma(t) = 0$: after t, continuation does not matter.

This is transition-dependent discounting, not necessarily a single global constant.

Theorem 2: Markov reward representation

Assumptions: Axioms 1-5 over lotteries over trajectories.

- Then, utility has a local recursive form:

$$u(t \cdot h) = r(t) + \gamma(t)u(h), \qquad u(\varepsilon) = 0.$$

Theorem 2: Markov reward representation

Assumptions: Axioms 1-5 over lotteries over trajectories.

- Then, utility has a local recursive form:

$$u(t \cdot h) = r(t) + \gamma(t)u(h), \qquad u(\varepsilon) = 0.$$

- Preferences over lotteries are still represented by expected utility.

Theorem 2: Markov reward representation

Assumptions: Axioms 1-5 over lotteries over trajectories.

- Then, utility has a local recursive form:

$$u(t \cdot h) = r(t) + \gamma(t)u(h), \qquad u(\varepsilon) = 0.$$

- Preferences over lotteries are still represented by expected utility.

This establishes a bridge from expected utility maximisation over whole-trajectory utility to RL-style expected sum of future reward maximisation.

Unrolling the recursion

- For $h = (t_1, t_2, \ldots, t_n)$:

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

Unrolling the recursion

- For $h = (t_1, t_2, \ldots, t_n)$:

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

- If $\gamma(t) \equiv \gamma$, this becomes standard discounted return:

$$u(h) = \sum_{k=1}^{n} \gamma^{k-1} r(t_k).$$

Summary: Preference, utility, and reward

Preference relation $\succcurlyeq$
ordering over trajectories or lotteries

↓

Ordinal utility
numerical labels for deterministic rankings

↓

(vNM) Expected utility
expected-utility scale for lotteries

↓

Reward and discount
local coordinate system for trajectory utility

Open questions regarding AI alignment

- Will powerful AIs necessarily become expected utility maximisers?

Open questions regarding AI alignment

- Will powerful AIs necessarily become expected utility maximisers?
- Instead of 'given a reward, how to optimise it' we could ask 'what kind of properties we want our reward to satisfy'. For example, are any of these properties (or their negation) useful for fostering:
  - Corrigibility? (e.g. incompleteness)

Open questions regarding AI alignment

- Will powerful AIs necessarily become expected utility maximisers?
- Instead of 'given a reward, how to optimise it' we could ask 'what kind of properties we want our reward to satisfy'. For example, are any of these properties (or their negation) useful for fostering:
  - Corrigibility? (e.g. incompleteness)
  - Robust values? (e.g. continuity)

Open questions regarding AI alignment

- Will powerful AIs necessarily become expected utility maximisers?
- Instead of 'given a reward, how to optimise it' we could ask 'what kind of properties we want our reward to satisfy'. For example, are any of these properties (or their negation) useful for fostering:
  - Corrigibility? (e.g. incompleteness)
  - Robust values? (e.g. continuity)
  - Something else?
- Is RLHF providing an ordinal or cardinal utility signal? Are there implications of this?

Open questions regarding AI alignment

- Will powerful AIs necessarily become expected utility maximisers?
- Instead of 'given a reward, how to optimise it' we could ask 'what kind of properties we want our reward to satisfy'. For example, are any of these properties (or their negation) useful for fostering:
  - Corrigibility? (e.g. incompleteness)
  - Robust values? (e.g. continuity)
  - Something else?
- Is RLHF providing an ordinal or cardinal utility signal? Are there implications of this?