Representation and Invariance in Reinforcement Learning

Samuel Allen Alexander[1][0000−0002−7930−110X]
Arthur Paul Pedersen[2][0000−0002−2164−6404]

[1]Independent Researcher, samuelallenalexander@gmail.com
[2]The City University of New York

Abstract. Researchers have formalized reinforcement learning (RL) in different ways. If an agent in one RL framework is to run within another RL framework's environments, the agent must first be converted, or mapped, into that other framework. In this paper, we lay foundations for studying relative-intelligence-preserving mappability between RL frameworks. We introduce a criterion which is sufficient for relative intelligence to be preserved according to one particular method of measuring intelligence. We show that this criterion cannot be met when mapping between certain deterministic and stochastic RL frameworks, suggesting inherent fundamental differences between these different versions of RL.

1 Introduction

If we changed the rules, would the wise become fools? In reinforcement learning (RL), agents and environments interact. The agent's objective is to learn to act in its environment in order to maximize its rewards. When an agent interacts with an environment, the agent and the environment take turns. On the agent's turn, the agent chooses an **action** (or a probability distribution over a set of actions, according to which an action is randomly selected). The action so chosen is thereupon transmitted to agent and environment. On the environment's turn, the environment chooses a **percept** to send to the agent in response (or a probability distribution over a set of percepts, according to which a percept is randomly selected). The percept so chosen is likewise transmitted to agent and environment. Each percept includes a numerical **reward** and an **observation**.

This is all simple enough, but there are many different ways to formally represent RL. These different representations can be organized according to answers to key questions, such as who goes first, what actions are permitted, what observations are allowed, and how numerical rewards are issued. Implicit in treatments of RL is that answers to these questions are inconsequential. The problem addressed in this paper is whether this is really so. If answers to such questions are inconsequential to problems for reinforcement learning, then evaluation of agent performance — measures of their relative intelligence — would be expected to be invariant with respect to transformations between different RL frameworks.

The present paper develops techniques for understanding the extent to which scales for measuring agent intelligence are invariant to different RL representations. To be clear, this is a very complicated subject, and the present paper should be considered to be an initial step in a thousand-mile journey. Even the problem of merely measuring RL agent intelligence in some fixed RL framework is quite difficult. How much more difficult is the problem of testing preservation of relative intelligence when agents are converted from one RL framework to another? This is important because researchers have been treating RL as if the details are irrelevant, speaking of RL as if there is some core approach in common that everyone agrees on, whereas in reality this is far from the case.

As a simple illustrative example, if someone proposed an RL framework where only the reward 0 were allowed, clearly that framework would be weaker than the frameworks used in practice. Yet it is not so clear whether the same would hold if the proposed RL framework allowed only the rewards $\{0, 1\}$. What if it allowed only the rewards $\{-1, 0, 1\}$? What if it allowed only natural number rewards? What if it allowed only rational number rewards? What if it allowed arbitrary real-valued rewards? Would all these frameworks be equivalent, even though none are equivalent to the 0-only-reward framework? What does this question even mean, and how would one even begin to answer it?

The above questions, and others like them, are the high-level considerations which motivated this paper. But we found the problem so overwhelmingly complicated that, after exploring the topic for years, we finally decided to limit this initial paper to a modest first stab at it. In this paper we will consider only four specific concrete RL frameworks. These four frameworks are mutually identical except for two parameters: whether agents are deterministic or stochastic; and whether environments are deterministic or stochastic. The question of which of these frameworks are equivalent to which, is already important, because all four types of RL are routinely used in practice (often with deterministic agents or environments masquerading as stochastic through the use of pseudo-random number generators), whereas a majority of the theoretical literature assumes both agent and environment are stochastic. It would therefore be scandalous if these four frameworks were not all equivalent. As a matter of fact, our analysis suggests they might be significantly non-equivalent.

We will introduce a notion of transformation from one RL framework to another, and we will show that, if relative intelligence is compared as suggested in [1], then such transformations preserve relative intelligence of agents. Thus, at least in some sense, the existence of such a transformation is a sufficient condition for one RL framework to be reducible to another—and if two RL frameworks are mutually reducible to each other, they are in that sense equivalent to each other.

2 Preliminaries

The following definition attempts to explicitly recognise a decision implicit in much of the RL literature.

Definition 1. *By a **reinforcement learning framework** (or **RL framework**) we mean a triple (A, E, V) where:*

1. *A is a set whose members are called **agents**;*
2. *E is a set whose members are called **environments**;*
3. *$V : A \times E \to \mathbb{R}$ is a function assigning to every agent $\pi \in A$ and environment $\mu \in E$ a **total expected reward** $V^\pi_\mu \in \mathbb{R}$ representing how well π performs in μ.*

While Definition 1 is not intended to be as general as possible, it encompasses standard variants of RL[1].

What does it mean for one RL framework to be reducible to another? Imagine you have access to agents which a laboratory designed for RL framework $\mathscr{F}$, but the environments you are interested in were designed for RL framework $\mathscr{F}'$. The agents designed for framework $\mathscr{F}$ can only run in environments designed for framework $\mathscr{F}$, so you cannot directly use them. You must somehow convert them to run in framework $\mathscr{F}'$. So for every agent π designed for framework $\mathscr{F}$, you need to transform it into an agent π^* designed for framework $\mathscr{F}'$. This transformation should be faithful in some sense, but what does that mean? This question is vague, but we can at least say one thing: the transformation should preserve relative performance. If π is better than ρ in framework $\mathscr{F}$, then π^* should be better than ρ^* in framework $\mathscr{F}'$. But π^* and ρ^* perform in framework $\mathscr{F}'$ environments, whereas π and ρ perform in framework $\mathscr{F}$ environments. And in this thought experiment, the environments you care about are in framework $\mathscr{F}'$. So for every such environment μ in framework $\mathscr{F}'$, the relative performance of π^* and ρ^* in μ should be compared with the relative performance of π and ρ, not in μ itself, as they are incompatible with μ, but rather with some appropriate transformation μ_* of μ, where μ_* is an environment in framework $\mathscr{F}$. This motivates the following definition.

Definition 2. *Suppose $\mathscr{F} = (A, E, V)$ and $\overline{\mathscr{F}} = (\overline{A}, \overline{E}, \overline{V})$ are RL frameworks. A **transformation** from $\mathscr{F}$ to $\overline{\mathscr{F}}$ is a pair $(\bullet^* : A \to \overline{A}, \bullet_* : \overline{E} \to E)$ of functions such that:*

1. *(Faithfulness) For all $\pi, \rho \in A$ and $\mu \in \overline{E}$, $\overline{V}^{\pi^*}_\mu < \overline{V}^{\rho^*}_\mu$ iff $V^\pi_{\mu_*} < V^\rho_{\mu_*}$.*
2. *(Nontriviality 1) There exist $\pi, \rho \in A$, $\mu \in \overline{E}$ such that $\overline{V}^{\pi^*}_\mu < \overline{V}^{\rho^*}_\mu$.*

[1] Subsumed variants include, for example, those in which (i) agents are deterministic while environments need not be (as in [15] or [22]), (ii) neither agent nor environment need be deterministic (as in [17]), (iii) rewards are multiplied by discount factors, as in [23], (iv) each percept also includes a true-or-false flag indicating whether or not the percept signals the start of a new "episode" (as in [11] or [23], (v) where rewards are restricted, e.g. to $\mathbb{Q}$ or (as in [17]) some finite subset of $\mathbb{Q}$, (vi) where available actions vary from turn to turn (as in [23] or [13]), (vii) where available actions vary from environment to environment (as in [11]), (viii) where environments are Markov decision processes (as in most of [23]), (ix) where the environment can secretly simulate the agent [3, 10, 7], (x) where, environments and/or agents must be computable.

3. (Nontriviality 2) There exist $\pi \in A$, $\mu, \nu \in \overline{E}$ *such that* $\overline{V}_{\mu}^{\pi^*} < \overline{V}_{\nu}^{\pi^*}$.

Example 1. Suppose $\mathscr{F}$ and $\overline{\mathscr{F}}$ are two RL frameworks identical in every way except that $\mathscr{F}$ permits rewards from $\mathbb{Z}$ but $\overline{\mathscr{F}}$ only permits rewards from $2\mathbb{Z}$, the set of *even* integers. We believe anyone familiar with RL would informally consider these two RL frameworks to be equivalent. The obvious transformation from $\mathscr{F}$ to $\overline{\mathscr{F}}$ is the pair $(\bullet^*, \bullet_*)$ defined as follows. For any agent π of $\mathscr{F}$, let π^* be the agent of $\overline{\mathscr{F}}$ which results from wrapping π with an intermediary function that divides all rewards by 2. And for any environment μ of $\overline{\mathscr{F}}$, let μ_* be the environment of $\mathscr{F}$ which takes an input, multiplies all the rewards in that input by 2, passes the mutated input to μ, and returns μ's output but with reward divided by 2. Similarly, a transformation from $\overline{\mathscr{F}}$ to $\mathscr{F}$ can be obtained by replacing 2 with $\frac{1}{2}$ above.

Lemma 1 (Composability).

Suppose $\mathscr{F}, \overline{\mathscr{F}}, \overline{\overline{\mathscr{F}}}$ *are RL frameworks. If there is a transformation from* $\mathscr{F}$ *to* $\overline{\mathscr{F}}$ *and a transformation from* $\overline{\mathscr{F}}$ *to* $\overline{\overline{\mathscr{F}}}$*, then there is a transformation from* $\mathscr{F}$ *to* $\overline{\overline{\mathscr{F}}}$.

Proof. Write $\mathscr{F} = (A, E, V)$, $\overline{\mathscr{F}} = (\overline{A}, \overline{E}, \overline{V})$, $\overline{\overline{\mathscr{F}}} = (\overline{\overline{A}}, \overline{\overline{E}}, \overline{\overline{V}})$. Assume $(\bullet^*, \bullet_*)$ is a transformation from $\mathscr{F}$ to $\overline{\mathscr{F}}$, so $\bullet^* : A \to \overline{A}$ and $\bullet_* : \overline{E} \to E$. Assume $(\bullet^\dagger, \bullet_\dagger)$ is a transformation from $\overline{\mathscr{F}}$ to $\overline{\overline{\mathscr{F}}}$, so $\bullet^\dagger : \overline{A} \to \overline{\overline{A}}$ and $\bullet_\dagger : \overline{\overline{E}} \to \overline{E}$. Define $\bullet^\ddagger : A \to \overline{\overline{A}}$ by $\pi^\ddagger = (\pi^*)^\dagger$ and define $\bullet_\ddagger : \overline{\overline{E}} \to E$ by $\mu_\ddagger = (\mu_\dagger)_*$. It is straightforward to show that $(\bullet^\ddagger, \bullet_\ddagger)$ is a transformation from $\mathscr{F}$ to $\overline{\overline{\mathscr{F}}}$.

Lemma 2. *(Self-reducibility) Suppose* $\mathscr{F}$ *is an RL framework. If* $\mathscr{F}$ *is nontrivial, in the sense that* $\mathscr{F}$ *contains agents* π, ρ, σ *and environments* μ, ν, τ *such that* $V_\mu^\pi < V_\mu^\rho$ *and* $V_\nu^\sigma < V_\tau^\sigma$*, then there is a translation from* $\mathscr{F}$ *to itself.*

Proof. Write $\mathscr{F} = (A, E, V)$. It is straightforward to show that $(\bullet^*, \bullet_*)$ is a translation from $\mathscr{F}$ to $\mathscr{F}$ where $\bullet^* : A \to A$ is the identify function on A and $\bullet_* : E \to E$ is the identity function on E.

3 Comparing intelligence using ultrafilters and preserving relative intelligence

The question of how to measure intelligence of RL agents is nontrivial, even in a given fixed RL framework. One proposal is the Legg-Hutter intelligence measure [17], but that proposed measure involves infinite sums and the noncomputable Kolmogorov complexity function, making the proposal mathematically unwieldy; we have not been able to prove intelligence preservation results in terms of Legg-Hutter intelligence. Instead, we will compare intelligence using an approach which is mathematically more tractable, originally introduced by [1].

The idea is that in order to compare two RL agents π and ρ, to determine which one is more intelligent (or whether they are equally intelligent), we can

consider these three possibilities to be candidates in an election, where environments are voters. For any particular environment μ, if $V_\mu^\pi > V_\mu^\rho$, then μ votes that π is more intelligent than ρ. If $V_\mu^\pi < V_\mu^\rho$, then μ votes that π is less intelligent than ρ. If $V_\mu^\pi = V_\mu^\rho$, then μ votes that π and ρ are equally intelligent. How can we decide the winner of such an election? It turns out there is an elegant way to do this using machinery from mathematical logic known as *ultrafilters* (we will motivate ultrafilters below, assuming no prior knowledge thereof).

3.1 Introduction to ultrafilters

We give an introduction to ultrafilters in terms of elections (they were previously introduced this way in [5] and in [4]). In this subsection, we fix a set E of environments. If the environments in E vote in an election between finitely many candidates, how can we determine which candidate wins?

Say that a subset $X \subseteq E$ is a **majority** if electoral victory would already be guaranteed given only the votes of X. Can we think of any axioms that majorities should satisfy?

Here are three fairly obvious axioms for majorities:

- (Properness) $\emptyset$ is not a majority (if no-one votes for you, you lose).
- (Monotonicity) If X is a majority and $Y \supseteq X$ then Y is a majority (additional votes should do no harm).
- (Maximality) If X is not a majority, then its complement X^c is a majority (in a two-candidate election, if one candidate does not win, then the other candidate wins).

A fourth axiom is much less obvious, and in fact is highly counter-intuitive if we rely on our intuition about *finite-voter* elections. We would probably never think of this next axiom if we were only thinking in terms of elections, but recall that we are particularly interested in a special type of election, namely, an intelligence-comparison election. We would very much like for the resulting agent comparator to be transitive. In other words, consider RL agents π, ρ, σ. If the voters vote that π is more intelligent than ρ, and also they vote that ρ is more intelligent than σ, then we would very much desire that they should vote that π is more intelligent than σ. To say the voters vote π more intelligent than ρ is to say that some majority X votes as much, and to say that they vote ρ more intelligent than σ is to say that some majority Y votes as much. Assuming *individual* voters are consistent, it would follow that $X \cap Y$ vote π more intelligent than σ. Thus, in order to achieve the desired transitivity, we enforce the following counter-intuitive axiom.

- ($\cap$-closure) If X and Y are majorities, then $X \cap Y$ is a majority.

It turns out that through these electoral considerations we have already arrived at the mathematically sophisticated notion of the ultrafilter.

Definition 3. *Suppose E is a set. By an **ultrafilter on** E we mean a set $\mathcal{U}$ of subsets of E (intuitively thought of as* majorities*) which satisfy Properness, Monotonicity, Maximality and $\cap$-closure.*

Thus, if the environments in the set E are going to vote in an election with finitely many candidates, one way to determine the winner is to fix an ultrafilter $\mathcal{U}$ on E and declare that for each candidate c, if $\{\mu \in E : \mu \text{ votes for } c\} \in \mathcal{U}$, then c wins. The $\cap$-closure and Properness axioms ensure at most one candidate can win. The Maximality axiom (possibly iterated if there are > 2 candidates) ensures at least one candidate must win. Economists have shown [16] that if we impose certain requirements on election-decision methods, then conversely, every election-decision method satisfying those requirements is one of these ultrafilter-based decision methods[2].

3.2 Preservation of relative intelligence by transformation of RL frameworks

The previous subsection motivates the following notion of relative intelligence.

Definition 4. *Suppose $\mathscr{F} = (A, E, V)$ is an RL framework. Let $\mathcal{U}$ be an ultrafilter on E. We define the intelligence comparator $\leq_{\mathcal{U}}$, a binary relation on A, as follows. For all $\pi, \rho \in A$, we declare $\pi \leq_{\mathcal{U}} \rho$ iff $\{\mu \in E : V_\mu^\pi \leq V_\mu^\rho\} \in \mathcal{U}$.*

In plain English: $\pi \leq_{\mathcal{U}} \rho$ if the environments *vote* that ρ performs at least as well as π (when we use $\mathcal{U}$ to decide the outcome of the election once the votes are cast). We leave the proof of the following lemma as an exercise to the reader (using the ultrafilter axioms, Definition 3).

Lemma 3. *Suppose $\mathscr{F} = (A, E, V)$ is an RL framework and $\mathcal{U}$ is an ultrafilter on E.*

1. *(Reflexivity) For every $\pi \in A$, $\pi \leq_{\mathcal{U}} \pi$.*
2. *(Transitivity) For all $\pi, \rho, \sigma \in A$, if $\pi \leq_{\mathcal{U}} \rho$ and $\rho \leq_{\mathcal{U}} \sigma$, then $\pi \leq_{\mathcal{U}} \sigma$.*

Since we are interested in preservation (or lack thereof) of relative intelligence by a transformation from one RL framework to another, we would like a way to transform the above relative intelligence notion between frameworks.

Definition 5. *(Transformation of an ultrafilter) Suppose $\mathscr{F} = (A, E, V)$ and $\overline{\mathscr{F}} = (\overline{A}, \overline{E}, \overline{V})$ are RL frameworks, $(\pi \mapsto \pi^* : A \to \overline{A}, \mu \mapsto \mu_* : \overline{E} \to E)$ is a transformation from $\mathscr{F}$ to $\overline{\mathscr{F}}$, and $\mathcal{U}$ is an ultrafilter on $\overline{E}$. For each $Y \subseteq \overline{E}$, let $Y_* = \{\mu_* : \mu \in Y\} \subseteq E$. We define*

$$\mathcal{U}_* = \{X \subseteq E : X \supseteq Y_* \text{ for some } Y \in \mathcal{U}\}.$$

[2] The requirements in question are exactly the desiderata from Arrow's Impossibility Theorem, minus non-dictatorialness. Non-dictatorial decision-methods correspond exactly with so-called *free ultrafilters*: an ultrafilter $\mathcal{U}$ on E is **free** if it has the property that there does not exist any $\mu \in E$ such that $\{\mu\} \in \mathcal{U}$. Assuming E is infinite, it is known that free ultrafilters on E exist. This does not contradict Arrow's Impossibility Theorem, because Arrow's Impossibility Theorem requires that the set of voters is finite.

Lemma 4. *For all $\mathscr{F}$, $\overline{\mathscr{F}}$, $(\bullet^*, \bullet_*)$, $\mathcal{U}$ as in Definition 5, $\mathcal{U}_*$ is an ultrafilter on E.*

Proof. Straightforward.

The following is a relative intelligence preservation theorem for RL framework transformations, in the following sense. It says that if we have a transformation from a source framework to a destination framework, and if we compare relative intelligence in the destination framework using the electoral method (deciding elections with some ultrafilter $\mathcal{U}$ on the destination framework's environments), then those comparisons are preserved by the transformation (if we decide elections with $\mathcal{U}_*$ on the source framework's environments).

Theorem 1. *(Preservation Theorem) Suppose $\mathscr{F} = (A, E, V)$, $\overline{\mathscr{F}} = (\overline{A}, \overline{E}, \overline{V})$ are RL frameworks and $(\bullet^* : A \to \overline{A}, \bullet_* : \overline{E} \to E)$ is a transformation from $\mathscr{F}$ to $\overline{\mathscr{F}}$. For any ultrafilter $\mathcal{U}$ on $\overline{E}$, the transformation $(\bullet^*, \bullet_*)$ preserves relative intelligence in the following sense: for all $\pi, \rho \in A$, we have $\pi \leq_{\mathcal{U}_*} \rho$ iff $\pi^* \leq_{\mathcal{U}} \rho^*$.*

Proof. ($\Leftarrow$) Assume $\pi^* \leq_{\mathcal{U}} \rho^*$. By definition, this means $Y \in \mathcal{U}$, where $Y = \{\mu \in \overline{E} : \overline{V}_\mu^{\pi^*} \leq \overline{V}_\mu^{\rho^*}\}$. Let $X = \{\nu \in E : V_\nu^\pi \leq V_\nu^\rho\}$, we must show $X \in \mathcal{U}_*$. By the Monotonicity property of ultrafilters, it suffices to show some subset of X is in $\mathcal{U}_*$. Since $Y \in \mathcal{U}$, it follows that $Y_* \in \mathcal{U}_*$; we will show $Y_* \subseteq X$. Compute:

$$\begin{aligned}
Y_* &= \{\mu_* : \mu \in Y\} && \text{(Def. of } Y_*\text{)} \\
&= \{\nu \in E : \nu = \mu_* \text{ for some } \mu \in Y\} && \text{(Rewriting)} \\
&= \{\nu \in E : \nu = \mu_* \text{ for some } \mu \in \overline{E} \text{ with } \overline{V}_\mu^{\pi^*} \leq \overline{V}_\mu^{\rho^*}\} && \text{(Def. of } Y\text{)} \\
&= \{\nu \in E : \nu = \mu_* \text{ for some } \mu \in \overline{E} \text{ with } V_{\mu_*}^\pi \leq V_{\mu_*}^\rho\} && \text{(Def. 2 part 1)} \\
&\subseteq \{\nu \in E : V_\nu^\pi \leq V_\nu^\rho\} \\
&= X. && \text{(Def. of } X\text{)}
\end{aligned}$$

($\Rightarrow$) By rewriting our proof of ($\Leftarrow$) with $\leq$ changed to $\not\leq$ throughout, we get a proof that if $\pi^* \not\leq_{\mathcal{U}} \rho^*$ then $\pi \not\leq_{\mathcal{U}_*} \rho$.

Remark 1. Readers interested in measurement theory will be interested in the following variation. Suppose $\mathscr{F} = (A, E, V)$, $\overline{\mathscr{F}} = (\overline{A}, \overline{E}, \overline{V})$ are RL frameworks and $(\bullet^* : A \to \overline{A}, \bullet_* : \overline{E} \to E)$ is a transformation from $\mathscr{F}$ to $\overline{\mathscr{F}}$. Call $(\bullet^*, \bullet_*)$ a **scaling transformation** if it satisfies the following additional property:

- For all $\pi, \rho \in A$, for all $\mu \in \overline{E}$, for all $k \in \mathbb{R}$, $\overline{V}_\mu^{\pi^*} < k\overline{V}_\mu^{\rho^*}$ iff $V_{\mu_*}^\pi < kV_{\mu_*}^\rho$.

Fix an ultrafilter $\mathcal{U}$ on $\overline{E}$. For all $\pi, \rho \in A$ and $k \in \mathbb{R}$, define:

- $\pi^* \leq_{\mathcal{U}} k\rho^*$ iff $\{\mu \in \overline{E} : \overline{V}_\mu^{\pi^*} \leq k\overline{V}_\mu^{\rho^*}\} \in \mathcal{U}$;
- $\pi \leq_{\mathcal{U}_*} k\rho$ iff $\{\nu \in E : V_\nu^\pi \leq kV_\nu^\rho\} \in \mathcal{U}_*$.

Then by almost identical reasoning to the proof of Theorem 1, one can show that for any scaling transformation $(\bullet^*, \bullet_*)$, $\pi \leq_{\mathcal{U}_*} k\rho$ iff $\pi^* \leq_{\mathcal{U}} k\rho^*$. Thus, scaling transformations preserve relative intelligence even more strongly: they preserve real ratio relations such as "π is at least twice as intelligent as ρ" or "π is not at least half as intelligent as ρ".

4 Concrete results

In this section we will introduce four specific concrete RL frameworks. This will involve fixing action-sets and percept-sets, defining histories, and defining specific performance measures V_μ^π. Bear in mind that this infrastructure is specific to the four specific concrete RL frameworks.

Fix nonempty finite sets $\mathcal{A}$, $\mathcal{E}$. Elements of $\mathcal{A}$ are **actions** and elements of $\mathcal{E}$ are **percepts**. We assume $\mathcal{A} \cap \mathcal{E} = \emptyset$. We typically write a member of $\mathcal{A}$ as x and a member $\mathcal{E}$ as y. Assume a function $R : \mathcal{E} \to \mathbb{Z}$ assigning[3] to every percept $x \in \mathcal{E}$ an integer-valued **reward**. We assume the range of R includes 0 and 1.

The fact that $R(x)$ is integer-valued is critical for some of the proofs below. We do not currently know whether the theorems in question would remain true if $R(x)$ were allowed to be an arbitrary element of $\mathbb{Q}$, which would be more relevant in practice: rewards in reinforcement learning are not usually restricted to be integer-valued (though some important environments do have integer-valued rewards, for example environments where the agent gets reward $+1$ for winning a game, -1 for losing a game, and 0 for any other move).

We define **histories** inductively so that: (i) The empty sequence is a history; (ii) for any percept x, $\langle x \rangle$ is a history; (iii) for any nonempty history h ending with a percept, for any action y, hy is a history (where hy is the result of appending y to h); (iv) for any nonempty history h ending with an action, for any percept x, hx is a history (where hx is the result of appending x to h). In plain English: a history is a finite sequence starting with a percept, followed by an action, followed by a percent, followed by an action, and so on for some finite number of steps (the empty sequence is also considered a history). An **agent history** is a history which ends with a percept (so named because these are the histories intended to be seen by the agent). An **environment history** is a history which is either empty or ends with an action. We write H_A for the set of all agent histories, and we write H_E for the set of all environment histories.

Lemma 5. *A history is an agent history iff it has odd length; it is an environment history iff it has even length.*

Proof. By induction.

For nonempty finite set X, let $\Delta(X)$ be the set of $\mathbb{Q}$-valued probability distributions on X.

Definition 6 (Deterministic and Stochastic Agents and Environments). *Define the following sets:*

$A^{\text{det}} = \mathcal{A}^{H_A}$*, the set of all functions* $\pi : H_A \to \mathcal{A}$ *(call these functions* ***deterministic agents****).*

[3] By abstracting the function R out, rather than requiring that percepts be observation-reward pairs, we simplify certain technical details. A similar device is used in [12]. See also [18] where two different versions of RL are implemented, one with observation-reward pairs, one with reward-observation pairs, in order to test whether some empirical results depend on the ordering of the pairs.

$E^{\text{det}} = \mathcal{E}^{H_E}$, *the set of all functions* $\mu : H_E \to \mathcal{E}$ *(call these functions* ***deterministic environments****).*
$A^{\text{rnd}} = \Delta(\mathcal{A})^{H_A}$, *the set of all functions* $\pi : H_A \to \Delta(\mathcal{A})$ *(call these functions* ***stochastic agents****).*
$E^{\text{rnd}} = \Delta(\mathcal{E})^{H_E}$, *the set of all functions* $\mu : H_E \to \Delta(\mathcal{E})$ *(call these functions* ***stochastic environments****).*

An environment provides a way to obtain a percept from an environment history. Appending that percept to that history yields an agent history. An agent then provides a way to obtain an action from an agent history, and appending that action to that agent history gives us another environment history, and the process can be repeated forever.

Definition 7 (Expected total reward). *Let* $\pi \in A^{\text{det}} \cup A^{\text{rnd}}$, $\mu \in E^{\text{det}} \cup E^{\text{rnd}}$.

1. *For every* $n \in \mathbb{N}$, *let* $V^{\pi}_{\mu,n}$ *be the expected value of the total reward-sum* $R(x_1) + \cdots + R(x_n)$ *if the history* $x_1 y_1 \ldots x_n y_n$ *is generated as follows:*
   - *If* $\mu \in E^{\text{det}}$ *then* $x_1 = \mu(\langle\rangle)$. *If* $\mu \in E^{\text{rnd}}$ *then* x_1 *is randomly chosen from* $\mathcal{E}$ *based on the probability distribution* $\mu(\langle\rangle) \in \Delta(\mathcal{E})$.
   - *If* $\pi \in A^{\text{det}}$ *then* $y_1 = \pi(\langle x_1 \rangle)$. *If* $\pi \in A^{\text{rnd}}$ *then* y_1 *is randomly chosen from* $\mathcal{A}$ *based on the probability distribution* $\pi(\langle x_1 \rangle) \in \Delta(\mathcal{A})$.
   - *For* $1 < i < n$, *if* $\mu \in E^{\text{det}}$ *then* $x_{i+1} = \mu(x_1 y_1 \ldots x_i y_i)$; *if* $\mu \in E^{\text{rnd}}$ *then* x_{i+1} *is randomly chosen from* $\mathcal{E}$ *based on the probability distribution* $\mu(x_1 y_1 \ldots x_i y_i) \in \Delta(\mathcal{E})$.
   - *For* $1 < i \leq n$, *if* $\pi \in A^{\text{det}}$ *then* $y_i = \pi(x_1 y_1 \ldots x_{i-1} y_{i-1} x_i)$; *if* $\pi \in A^{\text{rnd}}$ *then* y_i *is randomly chosen from* $\mathcal{A}$ *based on the probability distribution* $\pi(x_1 y_1 \ldots x_{i-1} y_{i-1} x_i)$.
2. *Let* $V^{\pi}_{\mu} = \lim_{n\to\infty} V^{\pi}_{\mu,n}$, *provided the limit converges to a real number. If not, then* V^{π}_{μ} *is undefined.*

Since the above V^{π}_{μ} does not always converge, it is not directly suitable for Definition 1. To get around this, we restrict our attention to environments μ for which V^{π}_{μ} always converges (this trick was introduced in [6]).

Definition 8 (Well-behaved environments).

1. *We say* $\mu \in E^{\text{det}} \cup E^{\text{rnd}}$ *is* ***well-behaved*** *if it has the following property: for every* $\pi \in A^{\text{det}} \cup A^{\text{rnd}}$, V^{π}_{μ} *exists.*
2. *Let* W^{det} *denote the set of well-behaved deterministic environments and* W^{rnd} *the set of well-behaved stochastic environments.*

A word on notation might be helpful. In the notation V^{π}_{μ}, the superscript on V is used for the agent, and the subscript on V is used for the environment. In the same way, in the following definition, the superscript on $\mathscr{F}$ refers to agents, and the subscript on $\mathscr{F}$ refers to environments.

Definition 9 (Four Specific RL Frameworks).

- *The **standard RL framework with deterministic agents and environments** is the RL framework* $\mathscr{F}^{\mathrm{det}}_{\mathrm{det}} = (A^{\mathrm{det}}, W^{\mathrm{det}}, V)$.
- *The **standard RL framework with stochastic agents and deterministic environments** is the RL framework* $\mathscr{F}^{\mathrm{rnd}}_{\mathrm{det}} = (A^{\mathrm{rnd}}, W^{\mathrm{det}}, V)$.
- *The **standard RL framework with deterministic agents and stochastic environments** is the RL framework* $\mathscr{F}^{\mathrm{det}}_{\mathrm{rnd}} = (A^{\mathrm{det}}, W^{\mathrm{rnd}}, V)$.
- *The **standard RL framework with stochastic agents and environments** is the RL framework* $\mathscr{F}^{\mathrm{rnd}}_{\mathrm{rnd}} = (A^{\mathrm{rnd}}, W^{\mathrm{rnd}}, V)$.

The following theorem is the main result of this paper. For the four concrete frameworks of Definition 9, we answer the $4 \cdot (4-1) = 12$ transformation-existence questions.

Theorem 2. *For all* $\mathscr{G}, \mathscr{H} \in \{\mathscr{F}^{\mathrm{rnd}}_{\mathrm{rnd}}, \mathscr{F}^{\mathrm{rnd}}_{\mathrm{det}}, \mathscr{F}^{\mathrm{det}}_{\mathrm{rnd}}, \mathscr{F}^{\mathrm{det}}_{\mathrm{det}}\}$ *with* $\mathscr{G} \neq \mathscr{H}$*, there is a transformation from* $\mathscr{G}$ *to* $\mathscr{H}$ *iff* $\mathscr{G} = \mathscr{F}^{\mathrm{rnd}}_{\mathrm{det}}$ *or* $\mathscr{H} = \mathscr{F}^{\mathrm{det}}_{\mathrm{rnd}}$*. In other words: there is a transformation from* $\mathscr{G}$ *to* $\mathscr{H}$ *iff there is an arrow from* $\mathscr{G}$ *to* $\mathscr{H}$ *in Figure 1.*

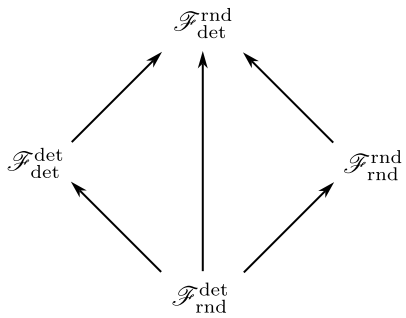

Fig. 1. Existence of transformations between four concrete RL frameworks.

We will prove Theorem 2 by a series of prelimary results.

4.1 Proof of Theorem 2

First, we will prove the positive parts of Theorem 2. We begin by defining embeddings of deterministic agents (resp. environments) among stochastic agents (resp. environments).

Lemma 6. *1. There exists a function* $\hat{\bullet} : A^{\mathrm{det}} \to A^{\mathrm{rnd}}$ *such that for all* $\mu \in W^{\mathrm{det}} \cup W^{\mathrm{rnd}}$, $V^{\pi}_{\mu} = V^{\hat{\pi}}_{\mu}$.

2. There exists a function $\hat{\bullet} : W^{\mathrm{det}} \to W^{\mathrm{rnd}}$ *such that for all* $\pi \in A^{\mathrm{det}} \cup A^{\mathrm{rnd}}$, $V^{\pi}_{\mu} = V^{\pi}_{\hat{\mu}}$.

Proof.
(1) Define $\hat{\pi} : H_A \to \Delta(\mathcal{A})$ by

$$\hat{\pi}(y|h) = \begin{cases} 1 & \text{if } y = \pi(h), \\ 0 & \text{otherwise.} \end{cases}$$

For any $\mu \in W^{\text{det}} \cup W^{\text{rnd}}$, by induction on n, it is easy to show that for each n, $V^{\hat{\pi}}_{\mu,n} = V^{\pi}_{\mu,n}$. Taking the limit as $n \to \infty$, we are done.
(2) Similar to (1), defining $\hat{\mu} : H_E \to \Delta(\mathcal{E})$ by

$$\hat{\mu}(x|h) = \begin{cases} 1 & \text{if } x = \mu(h), \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 1 (Positive parts of Theorem 2). *For every arrow in Figure 1 from a source RL framework to a destination RL framework, there is a transformation from said source framework to said destination framework.*

Proof. (From $\mathscr{F}^{\text{det}}_{\text{det}}$ to $\mathscr{F}^{\text{rnd}}_{\text{det}}$) Define $\bullet^* : A^{\text{det}} \to A^{\text{rnd}}$ by $\pi^* = \hat{\pi}$ and define $\bullet_* : W^{\text{det}} \to W^{\text{det}}$ by $\mu_* = \mu$, where $\hat{\pi}$ is as in Lemma 6. It is straightforward to show $(\bullet^*, \bullet_*)$ is a transformation from $\mathscr{F}^{\text{det}}_{\text{det}}$ to $\mathscr{F}^{\text{rnd}}_{\text{det}}$ (for Nontriviality 1 and Nontriviality 2 (Definition 2), use the fact the range of R includes 0 and 1).
(From $\mathscr{F}^{\text{det}}_{\text{rnd}}$ to $\mathscr{F}^{\text{det}}_{\text{det}}$) Define $\bullet^* : A^{\text{det}} \to A^{\text{det}}$ by $\pi^* = \pi$ and define $\bullet_* : W^{\text{rnd}} \to W_{\text{det}}$ by $\mu_* = \hat{\mu}$, where $\hat{\mu}$ is as in Lemma 6. It is straightforward to show $(\bullet^*, \bullet_*)$ is a transformation from $\mathscr{F}^{\text{det}}_{\text{rnd}}$ to $\mathscr{F}^{\text{det}}_{\text{det}}$.
The other three arrows are similar.

To prove the negative parts of Theorem 2, we will need to take mixtures of agents and environments.

Lemma 7 (Mixing Lemma).

1. *Given weights $w_1, \ldots, w_n \in (0,1) \cap \mathbb{Q}$, with $w_1 + \cdots + w_n = 1$, and agents $\pi_1, \ldots, \pi_n \in A^{\text{rnd}}$, there exists $\pi \in A^{\text{rnd}}$ such that for every $\mu \in W^{\text{det}} \cup W^{\text{rnd}}$, $V^{\pi}_{\mu} = w_1 V^{\pi_1}_{\mu} + \cdots + w_n V^{\pi_n}_{\mu}$.*
2. *Given weights $w_1, \ldots, w_n \in (0,1) \cap \mathbb{Q}$, with $w_1 + \cdots + w_n = 1$, and environments $\mu_1, \ldots, \mu_n \in W^{\text{rnd}}$, there exists $\mu \in W^{\text{rnd}}$ such that for all $\pi \in A^{\text{det}} \cup A^{\text{rnd}}$, $V^{\pi}_{\mu} = w_1 V^{\pi}_{\mu_1} + \cdots + w_n V^{\pi}_{\mu_n}$.*

Proof. (1) For every history h and every $\rho \in A^{\text{rnd}}$, let $P^{\rho}(h)$ be the probability that h would be an initial segment of the percept-action sequence that would be randomly generated if ρ interacted with some environment, subject to the condition that that environment initially outputs the percepts in h. As in [8], define $\pi : H_A \to \Delta(\mathcal{A})$ by

$$\pi(y|h) = \frac{w_1 P^{\pi_1}(hy) + \cdots + w_n P^{\pi_n}(hy)}{w_1 P^{\pi_1}(h) + \cdots + w_n P^{\pi_n}(h)}$$

provided the denominator is $\neq 0$, or $\pi(y|h) = 1/|\mathcal{A}|$ otherwise. By an inductive argument on k, the expected total reward $V^{\pi}_{\mu,k}$ which π would obtain after k steps interacting with any well-behaved environment μ equals $w_1 V^{\pi_1}_{\mu,k} + \cdots + w_n V^{\pi_n}_{\mu,k}$, the weighted average of the expected rewards $\pi_1, \ldots, \pi_n$ would obtain after k steps interacting with μ. Taking the limit as $k \to \infty$, we are done. For details, see [8].

(2) Similar to (1), with $\mu : H_E \to \Delta(\mathcal{E})$ defined as follows. For every history h and every $\nu \in W^{\mathrm{rnd}}$, let $P_\nu(h)$ be the probability that h would be an initial segment of the percept-action sequence that would be randomly generated if ν interacted with some agent, subject to the condition that that agent initially outputs the actions in h. Define

$$\mu(x|h) = \frac{w_1 P_{\mu_1}(hx) + \cdots + w_n P_{\mu_n}(hx)}{w_1 P_{\mu_1}(h) + \cdots + w_n P_{\mu_n}(h)}$$

provided the denominator is $\neq 0$, or $\mu(x|h) = 1/|\mathcal{E}|$ otherwise. For details, adapt the proof of Lemma 48 (part 4) of [8] to a finite vector of weights (the "strongly well-behaved" hypothesis of said lemma can be replaced by "well-behaved" because of the finiteness of the vector of weights).

Theorem 3. *There does not exist a transformation from $\mathscr{F}^{\mathrm{det}}_{\mathrm{det}}$ to $\mathscr{F}^{\mathrm{rnd}}_{\mathrm{rnd}}$.*

Proof. For sake of contradiction, assume $(\bullet^* : A^{\mathrm{det}} \to A^{\mathrm{rnd}}, \bullet_* : W^{\mathrm{rnd}} \to W^{\mathrm{det}})$ is a transformation. By Nontriviality 2 (Definition 2), pick $\pi \in A^{\mathrm{det}}$ and $\mu, \nu \in W^{\mathrm{rnd}}$ such that $V^{\pi^*}_{\mu} < V^{\pi^*}_{\nu}$. For every $\alpha \in (0,1) \cap \mathbb{Q}$, using Theorem 7 (part 2) (with $n = 2$, $w_1 = \alpha$, $w_2 = 1-\alpha$), let σ_α be the result of mixing μ and ν, giving α weight to μ and $1-\alpha$ weight to ν, so $V^{\pi^*}_{\sigma_\alpha} = \alpha V^{\pi^*}_{\mu} + (1-\alpha) V^{\pi^*}_{\nu}$. Thus for all such rational $\alpha < \beta$, we have $V^{\pi}_{(\sigma_\alpha)_*} < V^{\pi}_{(\sigma_\beta)_*}$. Choose rationals $\ell_1 < \ell_2 < \ell_3 < \cdots$ in $(0, \frac{1}{2}) \cap \mathbb{Q}$ and choose rational $r \in (\frac{1}{2}, 1) \cap \mathbb{Q}$, so $r > \ell_i$ for every i. For each i, let $\tau_i = \sigma_{\ell_i}$. We have $V^{\pi}_{(\tau_1)_*} < V^{\pi}_{(\tau_2)_*} < \cdots$, and yet $V^{\pi}_{(\sigma_r)_*} > V^{\pi}_{(\tau_i)_*}$ for every i. This is impossible since $V^{\pi}_{(\sigma_r)_*} \in \mathbb{Z}$ and each $V^{\pi}_{(\tau_i)_*} \in \mathbb{Z}$: there does not exist an infinite strictly ascending sequence of integers *and* another integer bigger than them all.

Theorem 4. *There does not exist a transformation from $\mathscr{F}^{\mathrm{rnd}}_{\mathrm{rnd}}$ to $\mathscr{F}^{\mathrm{det}}_{\mathrm{det}}$.*

Proof. For sake of contradiction, assume $(\bullet^* : A^{\mathrm{rnd}} \to A^{\mathrm{det}}, \bullet_* : W^{\mathrm{det}} \to W^{\mathrm{rnd}})$ is a transformation. By Nontriviality 1 (Definition 2), pick $\pi, \rho \in A^{\mathrm{rnd}}$ and $\mu \in W^{\mathrm{det}}$ such that $V^{\pi^*}_{\mu} < V^{\rho^*}_{\mu}$, so $V^{\pi}_{\mu_*} < V^{\rho}_{\mu_*}$. Using Theorem 7 (part 1) (with $n = 2$, $w_1 = \alpha$, $w_2 = 1-\alpha$), for every $\alpha \in (0,1) \cap \mathbb{Q}$, let σ_α be the result of mixing π and ρ, giving α weight to π and $1-\alpha$ weight to ρ. So $V^{\sigma_\alpha}_{\mu_*} = \alpha V^{\pi}_{\mu_*} + (1-\alpha) V^{\rho}_{\mu_*}$. Thus for all such rationals $\alpha < \beta$, we have $V^{\sigma_\alpha}_{\mu_*} < V^{\sigma_\beta}_{\mu_*}$, so $V^{\sigma^*_\alpha}_{\mu} < V^{\sigma^*_\beta}_{\mu}$. Choose $\ell_1 < \ell_2 < \cdots$ in $(0, \frac{1}{2}) \cap \mathbb{Q}$ and $r \in (\frac{1}{2}, 1) \cap \mathbb{Q}$, so $r > \ell_i$ for each i. For every i, let $\tau_i = \sigma_{\ell_i}$. We have $V^{\tau^*_1}_{\mu} < V^{\tau^*_2}_{\mu} < \cdots$, and yet $V^{\sigma^*_r}_{\mu} > V^{\tau^*_i}_{\mu}$ for each i. This is impossible for the same reason as in Theorem 3.

Theorem 5. *(Negative parts of Theorem 2) For all distinct RL frameworks $\mathscr{G}$ and $\mathscr{H}$ in Figure 1, if the figure does not include an arrow from $\mathscr{G}$ to $\mathscr{H}$, then there is no transformation from $\mathscr{G}$ to $\mathscr{H}$.*

Proof. (From $\mathscr{F}^{\text{det}}_{\text{det}}$ to $\mathscr{F}^{\text{rnd}}_{\text{rnd}}$ and from $\mathscr{F}^{\text{rnd}}_{\text{rnd}}$ to $\mathscr{F}^{\text{det}}_{\text{det}}$) By Theorems 3 and 4.
(From $\mathscr{F}^{\text{rnd}}_{\text{det}}$ to $\mathscr{F}^{\text{det}}_{\text{det}}$) By Theorem 1, there is a transformation from $\mathscr{F}^{\text{rnd}}_{\text{rnd}}$ to $\mathscr{F}^{\text{rnd}}_{\text{det}}$. If there were a transformation from $\mathscr{F}^{\text{rnd}}_{\text{det}}$ to $\mathscr{F}^{\text{det}}_{\text{det}}$, then by composability (Lemma 1), there would be a transformation from $\mathscr{F}^{\text{rnd}}_{\text{rnd}}$ to $\mathscr{F}^{\text{det}}_{\text{det}}$. This would violate the previous case.
(From $\mathscr{F}^{\text{det}}_{\text{det}}$ to $\mathscr{F}^{\text{det}}_{\text{rnd}}$) By Theorem 1, there is a transformation from $\mathscr{F}^{\text{det}}_{\text{rnd}}$ to $\mathscr{F}^{\text{rnd}}_{\text{rnd}}$. If there were a transformation from $\mathscr{F}^{\text{det}}_{\text{det}}$ to $\mathscr{F}^{\text{det}}_{\text{rnd}}$, then by composability (Lemma 1), there would be a transformation from $\mathscr{F}^{\text{det}}_{\text{det}}$ to $\mathscr{F}^{\text{rnd}}_{\text{rnd}}$. This would violate the first case.
The remaining three negative results are proved similarly.

Combining Proposition 1 and Theorem 5, we have proved Theorem 2.

Theorem 2 suggests that, at least if rewards are limited to integers, the nature of reinforcement learning may be inherently different depending whether agents be deterministic or stochastic, and whether environments be deterministic or stochastic. We do not currently know whether Theorem 2 would remain true if arbitrary rational-number rewards were allowed. For lack of any better evidence, though, the analysis here at least urges that researchers should exercise caution before speaking about RL as if these decisions don't matter.

The positive parts of Theorem 2 can be slightly strengthened: it can be shown that the transformations in the proof of Theorem 1 are in fact scaling transformations, in the sense of Remark 1.

5 Summary and conclusion

This paper is intended as a tentative initial step toward the difficult problem of comparing different reinforcement learning frameworks in general. Different authors all have the same high-level intuition about RL, but the formal details vary from author to author: which formalizations of RL are equivalent to each other, and which formalizations are fundamentally different? As an extreme example, a version of RL where rewards are required to always be 0, is *clearly* weaker than all ordinary versions of RL. But what does that formally even mean?

We introduced (Definition 2) the notion of a *transformation* from one RL framework to another. In Section 3 we recalled from [1] an elegant method of comparing RL agent intelligence based on electoral considerations. The high-level idea is to consider RL environments to be voters who vote (based on the performance of agents in those environments) to decide whether one agent is more intelligent than another (or whether both are equally intelligent). These intelligence-competition elections have to be decided somehow, and economists in the 1970s showed that the election decision procedures satisfying certain desiderata correspond exactly to so-called *ultrafilters*, which we recalled for the reader who might not be familiar with them (Subsection 3.1). We showed (Theorem 1)

that if a transformation exists from a source RL framework to a destination RL framework, then this transformation can be used to transform ultrafilter-based intelligence comparators in the destination framework into ultrafilter-based intelligence comparators in the source framework, in such a way as to preserve the relative intelligence of agents.

The reason for introducing transformations is that we intend them to be a proxy for the intuitive notion of one RL framework being reducible to another. Given two RL frameworks, if each is reducible to the other in this sense, then that serves as a proxy for the intuitive notion of the two frameworks being equivalent.

We introduce (Definition 9) four concrete RL frameworks, differing from each other only in terms of two binary parameters: (i) whether agents are deterministic or stochastic, and (ii) whether environments are deterministic or stochastic. This gives rise to $4 \cdot 3 = 12$ questions about existence of transformations. We answer all twelve questions, five positively and seven negatively (Theorem 2); no two of the four frameworks are equivalent in the sense of having transformations going in both directions. This is evidence suggesting that all four frameworks might be mutually non-equivalent.

Our high-level hope is that these results will encourage authors to be more specific, when talking about reinforcement learning, about which version of RL they mean. For example, when Silver et al suggest [22] that RL will lead to AGI, which version of RL do they mean? (Silver et al do claim to provide a definition in that paper, but their definition is not rigorous, for example it is unclear exactly which numbers rewards are allowed to be.) Furthermore, we hope that the question of existence of transformations from one RL framework to another will be a source of much interesting mathematics.

Our RL framework definition is quite general but not as general as possible. More extreme variations of RL will require, in future work, more general RL framework notions (but the ideas in this paper serve as a template for how one can explore intelligence preservation results in those more general frameworks)[4].

Acknowledgments

We acknowledge Cole Wyeth and Aram Ebtakar for comments and feedback.

References

1. Samuel Allen Alexander. Intelligence via ultrafilters: structural properties of some intelligence comparators of deterministic Legg-Hutter agents. *Journal of Artificial General Intelligence*, 10:24–45, 2019.

[4] Some such variations include, for example, (i) RL with multiple reward-signals, (ii) RL with multiple agents [20, 9, 24, 19, 26, 14], (iii) preference-based RL [25], (iv) RL with rewards from non-Archimedean number systems allowing infinitary or infinitesimal rewards (suggested by [2], implicitly suggested by [27], and conspicuously not ruled out by [22]), (v) along the same lines, RL involving non-Archimedean probabilities [21], (vi) RL where V_μ^π is allowed to diverge (and relative performance of agents in an environment is defined in various ways accordingly).

2. Samuel Allen Alexander. The Archimedean trap: Why traditional reinforcement learning will probably not yield AGI. *Journal of Artificial General Intelligence*, 11:70–85, 2020.
3. Samuel Allen Alexander, Michael Castaneda, Kevin Compher, and Oscar Martinez. Extending environments to measure self-reflection in reinforcement learning. *Journal of Artificial General Intelligence*, 2022.
4. Samuel Allen Alexander and Bryan Dawson. Big-oh notations, elections, and hyperreal numbers: A Socratic dialogue. *Proceedings of the ACMS*, 23:15–22, 2022.
5. Samuel Allen Alexander and Bill Hibbard. Measuring intelligence and growth rate: Variations on Hibbard's intelligence measure. *Journal of Artificial General Intelligence*, 12(1):1–25, 2021.
6. Samuel Allen Alexander and Marcus Hutter. Reward-punishment symmetric universal intelligence. *AGI*, 2021.
7. Samuel Allen Alexander and Arthur Paul Pedersen. Pseudo-visibility: A game mechanic involving willful ignorance. In *FLAIRS*, 2022.
8. Samuel Allen Alexander, David Quarel, Len Du, and Marcus Hutter. Universal agent mixtures and the geometry of intelligence. In *AISTATS*. PMLR, 2023.
9. Alan W Beggs. On the convergence of reinforcement learning. *Journal of Economic Theory*, 122(1):1–36, 2005.
10. James Henry Bell, Linda Linsefors, Caspar Oesterheld, and Joar Skalse. Reinforcement learning in Newcomblike environments. In *NeurIPS*, 2021.
11. Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. *Preprint*, 2016.
12. Frank Feys, Helle Hvid Hansen, and Lawrence S Moss. Long-term values in Markov decision processes, (co)algebraically. In *International Workshop on Coalgebraic Methods in Computer Science*, pages 78–99. Springer, 2018.
13. Matthew Hausknecht, Karthik Narasimhan, and Shunyu Yao. Building stronger semantic understanding into text game reinforcement learning agents. *Microsoft Research Blog*, 2021.
14. José Hernández-Orallo, David L Dowe, Sergio Espana-Cubillo, M Victoria Hernández-Lloreda, and Javier Insa-Cabrera. On more realistic environment distributions for defining, evaluating and developing intelligence. In *AGI*, 2011.
15. Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer, 2004.
16. Alan P Kirman and Dieter Sondermann. Arrow's theorem, many agents, and invisible dictators. *Journal of Economic Theory*, 5(2):267–277, 1972.
17. Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444, 2007.
18. Shane Legg and Joel Veness. An approximation of the universal intelligence measure. In *Algorithmic Probability and Friends: Bayesian Prediction and Artificial Intelligence*, pages 236–249. Springer, 2013.
19. Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *ICML*, 1994.
20. Michael L Littman and Csaba Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *ICML*, volume 96, pages 310–318, 1996.
21. Arthur Paul Pedersen. Comparative expectations. *Studia Logica*, 102:811–848, 2014.
22. David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 2021.
23. Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

24. Gerhard Weiss. Collective learning of action sequences. *International Conference on Distributed Computing Systems*, pages 203–209, 1993.
25. Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1):4945–4990, 2017.
26. Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*, pages 321–384. Springer International Publishing, Cham, 2021.
27. Yufan Zhao, Michael R Kosorok, and Donglin Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294–3315, 2009.