

# Symmetric Complexity: On Legg-Hutter Universal Intelligence with Punishments

Samuel Allen Alexander

March 13, 2021

## Abstract

Can an agent’s intelligence level be negative? We extend the Legg-Hutter agent-environment framework to include punishments and argue for an affirmative answer to that question. We show that if the background encodings and Universal Turing Machine (UTM) admit certain Kolmogorov complexity symmetries, then the resulting Legg-Hutter intelligence measure is symmetric about the origin. In particular, this implies reward-ignoring agents have Legg-Hutter intelligence 0 according to such UTMs.

## 1 Introduction

In their paper [7], Legg and Hutter write:

“As our goal is to produce a definition of intelligence that is as broad and encompassing as possible, the space of environments used in our definition should be as large as possible.”

So motivated, we investigate what would happen if we extended the universe of environments to include environments with rewards from  $\mathbb{Q} \cap [-1, 1]$  instead of just from  $\mathbb{Q} \cap [0, 1]$  as in Legg and Hutter’s paper. In other words, we investigate what would happen if environments are not only allowed to reward agents but also to punish agents (a punishment being a negative reward).

We discovered that when negative rewards are allowed, this introduces a certain algebraic structure into the agent-environment framework. The main objection we anticipate to our extended framework is that it implies the negative intelligence of certain agents<sup>1</sup>. We would argue that this makes perfect sense when environments are capable of punishing agents: the intelligence of a reinforcement learning agent should measure the degree to which that agent

---

<sup>1</sup>Thus, this paper falls under the broader program of advocating for intelligence measures having different ranges than the nonnegative reals. Alexander has advocated more extreme extensions of the range of intelligence measures [2] [3]; by contrast, here we merely question the assumption that intelligence never be negative, leaving aside the question of whether intelligence should be real-valued.

can extract large rewards on average across many environments; an agent who instead extracts large punishments on average across many environments should therefore have a negative intelligence level.

The structure of the paper is as follows:

- In Section 2, we give preliminary definitions.
- In Section 3, we introduce what we call the dual of an agent and of an environment, and prove some algebraic theorems about these.
- In Section 4, we show the existence of UTMs yielding Kolmogorov complexities with certain symmetries, and show that the resulting Legg-Hutter intelligence measures are symmetric too.
- In Section 5, we summarize and make concluding remarks, including remarks about how these ideas might be applied to certain other intelligence measures.

## 2 Preliminaries

In defining agent and environment below, we attempt to follow Legg and Hutter [7] as closely as possible, except that we permit environments to output rewards from  $\mathbb{Q} \cap [-1, 1]$  rather than just  $\mathbb{Q} \cap [0, 1]$  (and, accordingly, we modify which well-behaved environments to restrict our attention to).

Throughout the paper, we implicitly fix a finite set  $\mathcal{A}$  of *actions*, a finite set  $\mathcal{O}$  of *observations*, and a finite set  $\mathcal{R} \subseteq \mathbb{Q} \cap [-1, 1]$  of *rewards* (so each reward is a rational number between  $-1$  and  $1$  inclusive), with  $|\mathcal{A}| > 1$ ,  $|\mathcal{O}| > 0$ ,  $|\mathcal{R}| > 0$ . We assume that  $\mathcal{R}$  has the following property: whenever  $\mathcal{R}$  contains any reward  $r$ , then  $\mathcal{R}$  also contains  $-r$ . We assume  $\mathcal{A}$ ,  $\mathcal{O}$ , and  $\mathcal{R}$  are mutually disjoint (i.e., no reward is an action, no reward is an observation, and no action is an observation). By  $\langle \rangle$  we mean the empty sequence.

**Definition 1.** (*Agents, environments, etc.*)

1. By  $(\mathcal{O}\mathcal{R}\mathcal{A})^*$  we mean the set of all finite sequences starting with an observation, ending with an action, and following the pattern “observation, reward, action, ...”. We include  $\langle \rangle$  in this set.
2. By  $(\mathcal{O}\mathcal{R}\mathcal{A})^*\mathcal{O}\mathcal{R}$  we mean the set of all sequences of the form  $s \frown o \frown r$  where  $s \in (\mathcal{O}\mathcal{R}\mathcal{A})^*$ ,  $o \in \mathcal{O}$  and  $r \in \mathcal{R}$  ( $\frown$  denotes concatenation).
3. By an agent, we mean a function  $\pi$  which assigns to every sequence  $s \in (\mathcal{O}\mathcal{R}\mathcal{A})^*\mathcal{O}\mathcal{R}$  a  $\mathbb{Q}$ -valued probability measure, written  $\pi(\bullet|s)$ , on  $\mathcal{A}$ . For every such  $s$  and every  $a \in \mathcal{A}$ , we write  $\pi(a|s)$  for  $(\pi(\bullet|s))(a)$ . Intuitively,  $\pi(a|s)$  is the probability that agent  $\pi$  will take action  $a$  in response to history  $s$ .

4. By an environment, we mean a function  $\mu$  which assigns to every  $s \in (\mathcal{O}\mathcal{R}\mathcal{A})^*$  a  $\mathbb{Q}$ -valued probability measure, written  $\mu(\bullet|s)$ , on  $\mathcal{O} \times \mathcal{R}$ . For every such  $s$  and every  $(o, r) \in \mathcal{O} \times \mathcal{R}$ , we write  $\mu(o, r|s)$  for  $(\mu(\bullet|s))(o, r)$ . Intuitively,  $\mu(o, r|s)$  is the probability that environment  $\mu$  will issue observation  $o$  and reward  $r$  to the agent in response to history  $s$ .
5. If  $\pi$  is an agent,  $\mu$  is an environment, and  $n \in \mathbb{N}$ , we write  $V_{\mu, n}^\pi$  for the expected value of the sum of the rewards which would occur in the sequence  $(o_0, r_0, a_0, \dots, o_n, r_n, a_n)$  randomly generated as follows:
  - (a)  $(o_0, r_0) \in \mathcal{O} \times \mathcal{R}$  is chosen randomly based on the probability measure  $\mu(\bullet|\langle \rangle)$ .
  - (b)  $a_0 \in \mathcal{A}$  is chosen randomly based on the probability measure  $\pi(\bullet|o_0, r_0)$ .
  - (c) For each  $i > 0$ ,  $(o_i, r_i) \in \mathcal{O} \times \mathcal{R}$  is chosen randomly based on the probability measure  $\mu(\bullet|o_0, r_0, a_0, \dots, o_{i-1}, r_{i-1}, a_{i-1})$ .
  - (d) For each  $i > 0$ ,  $a_i \in \mathcal{A}$  is chosen randomly based on the probability measure  $\pi(\bullet|o_0, r_0, a_0, \dots, o_{i-1}, r_{i-1}, a_{i-1}, o_i, r_i)$ .
6. If  $\pi$  is an agent and  $\mu$  is an environment, let  $V_\mu^\pi = \lim_{n \rightarrow \infty} V_{\mu, n}^\pi$ . Intuitively,  $V_\mu^\pi$  is the expected total reward which  $\pi$  would extract from  $\mu$ .

Note that it is possible for  $V_\mu^\pi$  to be undefined. For example, if  $\mu$  is an environment which always issues reward  $(-1)^n$  in response to the agent's  $n$ th action, then  $V_\mu^\pi$  is undefined for every agent  $\pi$ . This would not be the case if rewards were required to be  $\geq 0$ , so this is one way in which allowing punishments complicates the resulting theory.

In order to define Legg-Hutter-style intelligence measures, it is, in any case, necessary to restrict attention to certain well-behaved environments. Legg and Hutter (in their Section 3.2) restrict attention to environments  $\mu$  that never give total reward more than 1, from which it follows that  $V_\mu^\pi \leq 1$ . This implication is less clear when environments can punish the agent, so we take a different approach: rather than limit the total reward the environment can give and use that to force  $V_\mu^\pi$  to be bounded, we will cut the middleman and directly assume that  $V_\mu^\pi$  is bounded.

**Definition 2.** An environment  $\mu$  is well-behaved if  $\mu$  is computable and the following condition holds: for every agent  $\pi$ ,  $V_\mu^\pi$  exists and  $-1 \leq V_\mu^\pi \leq 1$ .

Note that reward-space  $[0, 1]$  can be transformed into punishment-space  $[-1, 0]$  either by means of  $r \mapsto -r$  or by means of  $r \mapsto r - 1$ . An advantage of  $r \mapsto -r$  is that it preserves well-behavedness of environments (we will prove this below in Corollary 7), whereas  $r \mapsto r - 1$  does not. For example, an environment which always issues reward 0 (so is well-behaved) would become an environment which always issues reward  $-1$  (so is ill-behaved). The authors of an interesting paper on reinforcement learning suicide/death [9] indicate surprise that certain agents change their behavior when the reward-space  $[0, 1]$  is transformed into the punishment-space  $[-1, 0]$ , because said agents do not

normally change their behavior in response to positive linear reward transformations like  $r \mapsto r - 1$ ; said behavior-changes are less surprising when one views the reward transformation as being the result of  $r \mapsto -r$  instead of  $r \mapsto r - 1$ .

### 3 Dual Agents and Dual Environments

In the Introduction, we promised that by allowing environments to punish agents, we would reveal algebraic structure not otherwise present. The key to this additional structure is the following definition.

**Definition 3.** (*Dual Agents and Dual Environments*)

1. For any sequence  $s$ , let  $\bar{s}$  be the sequence obtained by replacing every reward  $r$  in  $s$  by  $-r$ .
2. Suppose  $\pi$  is an agent. We define a new agent  $\bar{\pi}$ , the dual of  $\pi$ , as follows: for each  $s \in (\mathcal{ORA})^* \mathcal{OR}$ , for each action  $a \in \mathcal{A}$ ,

$$\bar{\pi}(a|s) = \pi(a|\bar{s}).$$

3. Suppose  $\mu$  is an environment. We define a new environment  $\bar{\mu}$ , the dual of  $\mu$ , as follows: for each  $s \in (\mathcal{ORA})^*$ , for each observation  $o \in \mathcal{O}$  and reward  $r \in \mathcal{R}$ ,

$$\bar{\mu}(o, r|s) = \mu(o, -r|\bar{s}).$$

**Lemma 4.** (*Double Negation*)

1. For each sequence  $s$ ,  $\bar{\bar{s}} = s$ .
2. For each agent  $\pi$ ,  $\bar{\bar{\pi}} = \pi$ .
3. For each environment  $\mu$ ,  $\bar{\bar{\mu}} = \mu$ .

*Proof.* Follows from the fact that for every real number  $r$ ,  $- - r = r$ .  $\square$

**Theorem 5.** Suppose  $\mu$  is an environment and  $\pi$  is an agent. Then

$$V_{\bar{\mu}}^{\bar{\pi}} = -V_{\mu}^{\pi}$$

(and the left-hand side is defined if and only if the right-hand side is defined).

*Proof.* By Definition 1 part 6, it suffices to show that for each  $n \in \mathbb{N}$ ,  $V_{\bar{\mu}, n}^{\bar{\pi}} = -V_{\mu, n}^{\pi}$ . For that, it suffices to show that for every  $s \in ((\mathcal{ORA})^*) \cup ((\mathcal{ORA})^* \mathcal{OR})$ , the probability  $X$  of generating  $s$  using  $\pi$  and  $\mu$  (as in Definition 1 part 5) equals the probability  $X'$  of generating  $\bar{s}$  using  $\bar{\pi}$  and  $\bar{\mu}$ . We will show this by induction on the length of  $s$ .

Case 1:  $s$  is empty. Then  $X = X' = 1$ .

Case 2:  $s$  terminates with a reward. Then  $s = t \frown o \frown r$  for some  $t \in (\mathcal{ORA})^*$ . Let  $Y$  (resp.  $Y'$ ) be the probability of generating  $t$  (resp.  $\bar{t}$ ) using  $\pi$  and  $\mu$  (resp.  $\bar{\pi}$  and  $\bar{\mu}$ ). We reason:

$$\begin{aligned}
X &= \mu(o, r|t)Y && \text{(Definition of } X\text{)} \\
&= \mu(o, -r|\bar{t})Y && \text{(Lemma 4)} \\
&= \bar{\mu}(o, -r|\bar{t})Y && \text{(Definition of } \bar{\mu}\text{)} \\
&= \bar{\mu}(o, -r|\bar{t})Y' && \text{(By induction, } Y = Y'\text{)} \\
&= X'. && \text{(Definition of } X'\text{)}
\end{aligned}$$

Case 3:  $s$  terminates with an action. Then  $s = t \frown a$  for some  $t \in (\mathcal{ORA})^*\mathcal{OR}$ . Let  $Y$  (resp.  $Y'$ ) be the probability of generating  $t$  (resp.  $\bar{t}$ ) using  $\pi$  and  $\mu$  (resp.  $\bar{\pi}$  and  $\bar{\mu}$ ). We reason:

$$\begin{aligned}
X &= \pi(a|t)Y && \text{(Definition of } X\text{)} \\
&= \pi(a|\bar{t})Y && \text{(Lemma 4)} \\
&= \bar{\pi}(a|\bar{t})Y && \text{(Definition of } \bar{\pi}\text{)} \\
&= \bar{\pi}(a|\bar{t})Y' && \text{(By induction, } Y = Y'\text{)} \\
&= X'. && \text{(Definition of } X'\text{)}
\end{aligned}$$

□

**Corollary 6.** *For every agent  $\pi$  and environment  $\mu$ ,*

$$V_{\mu}^{\pi} = -V_{\mu}^{\bar{\pi}}$$

*(and the left-hand side is defined if and only if the right-hand side is defined).*

*Proof.* If neither side is defined, then there is nothing to prove. Assume the left-hand side is defined. Then

$$\begin{aligned}
V_{\mu}^{\pi} &= V_{\mu}^{\bar{\bar{\pi}}} && \text{(Lemma 4)} \\
&= -V_{\mu}^{\bar{\pi}}, && \text{(Theorem 5)}
\end{aligned}$$

as desired. A similar argument holds if we assume the right-hand side is defined.

□

**Corollary 7.** *For every environment  $\mu$ ,  $\mu$  is well-behaved if and only if  $\bar{\mu}$  is well-behaved.*

*Proof.* We prove the  $\Rightarrow$  direction, the other is similar. Since  $\mu$  is well-behaved,  $\mu$  is computable, so clearly  $\bar{\mu}$  is computable. Let  $\pi$  be any agent. Since  $\mu$  is well-behaved,  $V_{\mu}^{\pi}$  is defined and  $-1 \leq V_{\mu}^{\pi} \leq 1$ . By Corollary 6,  $V_{\mu}^{\pi} = -V_{\mu}^{\bar{\pi}}$  is defined, implying  $-1 \leq V_{\mu}^{\bar{\pi}} \leq 1$ . By arbitrariness of  $\pi$ , this shows  $\bar{\mu}$  is well-behaved. □

## 4 Symmetric Intelligence

By definition, agent  $\bar{\pi}$  acts exactly as agent  $\pi$  would act if  $\pi$  were confused into thinking that punishments were rewards and rewards were punishments. Whatever ingenuity  $\pi$  applies to maximize rewards,  $\bar{\pi}$  applies that exact same ingenuity to maximize punishments. Thus, if  $\Upsilon$  is an intelligence function (intended to measure how well an agent extracts rewards in some aggregate sense across the whole infinite space of all well-behaved environments), then  $\Upsilon$  ought to satisfy the equation:

$$\Upsilon(\bar{\pi}) = -\Upsilon(\pi).$$

Not every intelligence function satisfies the above equation, but the above equation should be considered desirable when we invent ways of measuring intelligence: all else being equal, an intelligence function which satisfies the above equation should be considered better than an intelligence function which does not<sup>2</sup>. In this section, we will show that the universal intelligence measure proposed by Legg and Hutter satisfies the above equation, provided a background UTM and encoding are suitably chosen.

We write  $2^*$  for the set of finite binary strings. We write  $f : \subseteq A \rightarrow B$  to indicate that  $f$  has codomain  $B$  and that  $f$ 's domain is some subset of  $A$ .

**Definition 8.** (*Prefix-free universal Turing machines*)

1. A partial computable function  $f : \subseteq 2^* \rightarrow 2^*$  is prefix-free if the following requirement holds:  $\forall p, p' \in 2^*$ , if  $p$  is a strict initial segment of  $p'$ , then  $f(p)$  and  $f(p')$  are not both defined.
2. A prefix-free universal Turing machine (or PFUTM) is a prefix-free partial computable function  $U : \subseteq 2^* \rightarrow 2^*$  such that the following condition holds. For every prefix-free partial computable function  $f : \subseteq 2^* \rightarrow 2^*$ ,  $\exists y \in 2^*$  such that  $\forall x \in 2^*$ ,  $f(x) = U(y \frown x)$ . In this case, we say  $y$  is a computer program for  $f$  in programming language  $U$ .

Legg-Hutter intelligence depends not only on a PFUTM but also on how the inputs and outputs of environments are encoded.

**Definition 9.** By an RL-encoding we mean a computable function  $\sqcap : (\mathcal{ORA})^* \cup M \rightarrow 2^*$  (where  $M$  is the set of  $\mathbb{Q}$ -valued probability-measures on  $\mathcal{O} \times \mathcal{R}$ ) such that the range of  $\sqcap$  is prefix-free (i.e., no element of its range is a strict initial

---

<sup>2</sup>This depends strongly on our assumption that when we measure intelligence we are measuring aggregate reward-extraction. This might clash with everyday ideas of intelligence: if a mischievous child answers every single question wrong on a true-false IQ test, we would say that child is clearly very intelligent, and is using said intelligence to deliberately fail the test. To quote Socrates, "Don't you think the ignorant person would often involuntarily tell the truth when he wished to say falsehoods, if it so happened, because he didn't know; whereas you, the wise person, if you should wish to lie, would always consistently lie?" [10] Likewise if an agent consistently manages to extract the worst possible rewards from many environments, we might think the agent is brilliant but masochistic. Nevertheless, we would not want to put said agent in charge of navigating environments for us (at least not without some reverse psychology).

segment of another). We say  $\sqcap$  is suffix-free if its range is suffix-free (i.e., no element of its range is a strict terminal segment of another). We write  $\lceil x \rceil$  for  $\sqcap(x)$ .

**Definition 10.** (Kolmogorov Complexity) Suppose  $U$  is a PFUTM and  $\sqcap$  is an RL-encoding.

1. For each computable environment  $\mu$ , the Kolmogorov complexity of  $\mu$  given by  $U, \sqcap$ , written  $K_U^\sqcap(\mu)$ , is the smallest  $n \in \mathbb{N}$  such that there is some computer program of length  $n$ , in programming language  $U$ , for the function  $f : \lceil (\mathcal{ORA})^* \rceil \rightarrow 2^*$  defined by  $f(\lceil s \rceil) = \lceil \mu(\bullet|s) \rceil$ .
2. We say  $U$  is symmetric in its  $\sqcap$ -encoded-environment cross-section (or simply that  $U$  is  $\sqcap$ -symmetric) if  $K_U^\sqcap(\mu) = K_U^\sqcap(\bar{\mu})$  for every computable environment  $\mu$ .

**Theorem 11.** For every suffix-free RL-encoding  $\sqcap$ , there exists a  $\sqcap$ -symmetric PFUTM.

*Proof.* Let  $U_0$  be a PFUTM, we will modify  $U_0$  to obtain a  $\sqcap$ -symmetric PFUTM. For readability's sake, write POS for 0 and NEG for 1. Thinking of  $U_0$  as a programming language, we define a new programming language  $U$  as follows. Every program in  $U$  must begin with one of the keywords POS or NEG. Outputs of  $U$  are defined as follows.

- $U(\text{POS} \frown x) = U_0(x)$ .
- To compute  $U(\text{NEG} \frown x)$ , find  $s \in (\mathcal{ORA})^*$  such that  $x = y \frown \lceil s \rceil$  for some  $y$  (if no such  $s$  exists, diverge). Note that  $s$  is unique by suffix-freeness of  $\sqcap$ . If  $U_0(y \frown \lceil \bar{s} \rceil) = \lceil m \rceil$  for some  $\mathbb{Q}$ -valued probability-measure  $m$  on  $\mathcal{O} \times \mathcal{R}$ , then let  $U(\text{NEG} \frown x) = \lceil \bar{m} \rceil$  where  $\bar{m}(o, r) = m(o, -r)$ . Otherwise, diverge.
- Informally: If  $x$  appears to be an instruction to plug  $s$  into computer program  $y$  to get a probability measure  $\mu(\bullet|s)$ , then instead plug  $\bar{s}$  into  $y$  and flip the resulting probability measure so that the output ends up being the flipped version of  $\mu(\bullet|\bar{s})$ , i.e.,  $\bar{\mu}(\bullet|s)$ .

By construction, whenever  $\text{POS} \frown y$  is a  $U$ -computer program for  $\lceil s \rceil \mapsto \lceil \mu(\bullet|s) \rceil$ ,  $\text{NEG} \frown y$  is an equal-length  $U$ -computer program for  $\lceil s \rceil \mapsto \lceil \bar{\mu}(\bullet|s) \rceil$ , and vice versa. It follows that  $U$  is  $\sqcap$ -symmetric.  $\square$

The proof of Theorem 11 proves more than required: any PFUTM can be modified to make a  $\sqcap$ -symmetric PFUTM if  $\sqcap$  is suffix-free. In some sense, the construction in the proof of Theorem 11 works by eliminating bias: reinforcement learning itself is implicitly biased in its convention that rewards be positive and punishments negative. We can imagine a pessimistic parallel universe where RL instead follows the opposite convention, and the RL in that parallel universe is no less valid than the RL in our own. To be unbiased in this sense, a computer program defining an environment should specify which of the two RL regimes it is operating in (hence the POS and NEG keywords).

**Definition 12.** Let  $W$  be the set of all well-behaved environments. Let  $\overline{W} = \{\overline{\mu} : \mu \in W\}$ .

**Lemma 13.**  $W = \overline{W}$ .

*Proof.* By Corollary 7. □

**Definition 14.** For every PFUTM  $U$ , RL-encoding  $\sqcap$ , and agent  $\pi$ , the Legg-Hutter universal intelligence of  $\pi$  given by  $U, \sqcap$ , written  $\Upsilon_U^\sqcap(\pi)$ , is

$$\Upsilon_U^\sqcap(\pi) = \sum_{\mu \in W} 2^{-K_U^\sqcap(\mu)} V_\mu^\pi.$$

The sum defining  $\Upsilon_U^\sqcap(\pi)$  is absolutely convergent because the summands in absolute value are dominated by the summands defining Chaitin's constant (this is why we insist on the UTM being prefix-free). Thus, by a well-known theorem from elementary calculus, the sum does not depend on which order the  $\mu \in W$  are enumerated.

Legg-Hutter intelligence has been accused of being subjective because of its UTM-sensitivity [8]. More optimistically, its UTM-sensitivity could be considered a feature, reflecting the fact that there are many kinds of intelligence. It could be used to measure intelligence in various contexts, by choosing UTMs appropriate for those contexts: one could even use it to measure, say, chess intelligence, by choosing a UTM where chess-related environments are easiest to program.

**Theorem 15.** (*Symmetry about the origin*) For every RL-encoding  $\sqcap$ , every  $\sqcap$ -symmetric PFUTM  $U$ , and agent  $\pi$ ,

$$\Upsilon_U^\sqcap(\overline{\pi}) = -\Upsilon_U^\sqcap(\pi).$$

*Proof.* Compute:

$$\begin{aligned} \Upsilon_U^\sqcap(\overline{\pi}) &= \sum_{\mu \in W} 2^{-K_U^\sqcap(\mu)} V_\mu^{\overline{\pi}} && \text{(Definition 14)} \\ &= - \sum_{\mu \in W} 2^{-K_U^\sqcap(\mu)} V_\mu^\pi && \text{(Corollary 6)} \\ &= - \sum_{\mu \in W} 2^{-K_U^\sqcap(\overline{\mu})} V_\mu^\pi && (U \text{ is } \sqcap\text{-symmetric}) \\ &= - \sum_{\mu \in \overline{W}} 2^{-K_U^\sqcap(\mu)} V_\mu^\pi && \text{(Change of variables)} \\ &= - \sum_{\mu \in W} 2^{-K_U^\sqcap(\mu)} V_\mu^\pi && \text{(Lemma 13)} \\ &= -\Upsilon_U^\sqcap(\pi). && \text{(Definition 14)} \end{aligned}$$

□



The above desideratum, that  $\Upsilon(\bar{\pi}) = -\Upsilon(\pi)$ , applies to numerical intelligence measures. If one is merely interested in binary intelligence comparators (such as those in [1]), the desideratum can be weakened into a non-numerical comparator form: If  $\pi$  is more intelligent than  $\rho$ , then  $\bar{\pi}$  should be less intelligent than  $\bar{\rho}$ . The following corollary addresses this desideratum.

**Corollary 16.** *For every RL-encoding  $\sqcap$ , every  $\sqcap$ -symmetric PFUTM  $U$ , for all agents  $\pi$  and  $\rho$ , if  $\Upsilon_U^\sqcap(\pi) > \Upsilon_U^\sqcap(\rho)$  then  $\Upsilon_U^\sqcap(\bar{\pi}) < \Upsilon_U^\sqcap(\bar{\rho})$ .*

*Proof.* By Theorem 15 and basic algebra.  $\square$

The following corollary shows that with suitable choice of encoding and UTM, Legg-Hutter universal intelligence satisfies another obvious desideratum.

**Corollary 17.** *Let  $\sqcap$  be an RL-encoding, let  $U$  be a  $\sqcap$ -symmetric PFUTM and suppose  $\pi$  is an agent which ignores rewards (by which we mean that  $\pi(\bullet|s)$  does not depend on the rewards in  $s$ ). Then  $\Upsilon_U^\sqcap(\pi) = 0$ .*

*Proof.* The hypothesis clearly implies  $\pi = \bar{\pi}$ . Thus by Theorem 15,  $\Upsilon_U^\sqcap(\pi) = -\Upsilon_U^\sqcap(\pi)$ , forcing  $\Upsilon_U^\sqcap(\pi) = 0$ .  $\square$

Corollary 17 illustrates why it is appropriate, for purposes of Legg-Hutter universal intelligence, to choose a  $\sqcap$ -symmetric PFUTM<sup>3</sup>. Consider an agent  $\pi_a$  which always blindly repeats a fixed action  $a \in \mathcal{A}$ . For any particular environment  $\mu$ , where  $\pi_a$  earns some total reward  $r$  by blind luck, that total reward ought to be cancelled out by  $\bar{\mu}$ , where the exact same blind luck becomes blind misfortune and  $\pi_a$  earns total reward  $-r$  (Corollary 6). But if  $K_U^\sqcap(\mu) \neq K_U^\sqcap(\bar{\mu})$ , then the different weights  $2^{-K_U^\sqcap(\mu)} \neq 2^{-K_U^\sqcap(\bar{\mu})}$  would prevent these two outcomes from cancelling each other.

We conclude this section with an exercise, suggesting how the techniques of this paper can be used to obtain other structural results.

**Exercise 18.** (*Permutations*)

1. For each permutation  $P : \mathcal{A} \rightarrow \mathcal{A}$  of the action-space, for each sequence  $s$ , let  $Ps$  be the result of applying  $P$  to all the actions in  $s$ . For each agent  $\pi$ , let  $P\pi$  be the agent defined by  $P\pi(a|s) = \pi(Pa|Ps)$ . For each environment  $\mu$ , let  $P\mu$  be the environment defined by  $P\mu(o, r|s) = \mu(o, r|Ps)$ . Show that in general  $V_\mu^\pi = V_{P^{-1}\mu}^{P\pi}$  and  $V_\mu^{P\pi} = V_{P\mu}^\pi$ .
2. Say that a PFUTM  $U$  is  $\sqcap$ -permutable if  $K_U^\sqcap(\mu) = K_U^\sqcap(P\mu)$  for every computable environment  $\mu$  and permutation  $P : \mathcal{A} \rightarrow \mathcal{A}$ . Show that if  $\sqcap$  is suffix-free then any given PFUTM can be transformed into a  $\sqcap$ -permutable PFUTM.
3. Show that if  $U$  is a  $\sqcap$ -permutable PFUTM, then  $\Upsilon_U^\sqcap(P\pi) = \Upsilon_U^\sqcap(\pi)$  for every agent  $\pi$  and permutation  $P : \mathcal{A} \rightarrow \mathcal{A}$ .
4. Modify this exercise to apply to permutations of the observation-space.

---

<sup>3</sup>Thus providing one answer to Leike and Hutter [8] who asked: “But what are other desirable properties of a UTM?”

## 5 Conclusion

By allowing environments to punish agents, we found additional algebraic structure in the agent-environment framework. Using this, we showed that certain Kolmogorov complexity symmetries yield Legg-Hutter intelligence symmetry.

In future work it would be interesting to explore how these symmetries manifest themselves in other Legg-Hutter-like intelligence measures [4] [5]. The precise strategy we employ in this paper is not relevant to Hibbardian sequence-prediction intelligence [6] [3]. However, a higher-level idea still applies: a predictor who intentionally mis-predicts should be considered less intelligent than a 0-intelligence blind guesser.

A high-level takeaway of this paper is that if intelligence is supposed to measure how much reward an agent extracts in aggregate across many environments, including environments capable of punishment, then such intelligence measures should sometimes output negative values.

## References

- [1] Samuel Allen Alexander. Intelligence via ultrafilters: structural properties of some intelligence comparators of deterministic Legg-Hutter agents. *JAGI*, 10(1):24–45, 2019.
- [2] Samuel Allen Alexander. The Archimedean trap: Why traditional reinforcement learning will probably not yield AGI. *JAGI*, 11(1):70–85, 2020.
- [3] Samuel Allen Alexander and Bill Hibbard. Measuring intelligence and growth rate: Variations on Hibbard’s intelligence measure. *JAGI*, 12(1):1–25, 2021.
- [4] Vaibhav Gavane. A measure of real-time intelligence. *JAGI*, 4(1):31–48, 2013.
- [5] José Hernández-Orallo and David L Dowe. Measuring universal intelligence: Towards an anytime intelligence test. *AI*, 174(18):1508–1539, 2010.
- [6] Bill Hibbard. Measuring agent intelligence via hierarchies of environments. In *ICAGI*, pages 303–308, 2011.
- [7] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444, 2007.
- [8] Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. In *Conference on Learning Theory*, pages 1244–1259. PMLR, 2015.
- [9] Jarryd Martin, Tom Everitt, and Marcus Hutter. Death and suicide in universal artificial intelligence. In *ICAGI*, 2016.
- [10] Plato. Lesser Hippias. In John M Cooper, Douglas S Hutchinson, et al., editors, *Plato: complete works*. Hackett Publishing, 1997.