Journal of Machine Learning Research 1 (2015) 1-9 Submitted 0/15; Published 0/15

On the Optimality of General Reinforcement Learners*

Jan Leike JAN.LEIKE@ANU.EDU.AU
Australian National University
Canberra, ACT, Australia

Marcus Hutter MARCUS.HUTTER@ANU.EDU.AU
Australian National University
Canberra, ACT, Australia

Editor:

Abstract

We discuss cases in which unlucky or adversarial choices of the universal prior cause the general reinforcement learning agent AIXI to misbehave drastically. We show that Legg-Hutter intelligence and thus balanced Pareto optimality is entirely subjective, and that every policy is Pareto optimal in the class of all computable environments. This undermines all existing optimality properties for AIXI. While it may still serve as a gold standard for general reinforcement learning, our results imply that AIXI is a *relative* theory, dependent on the choice of the universal prior.

Keywords: AIXI, general reinforcement learning, universal Turing machine, Legg-Hutter intelligence, balanced Pareto optimality, asymptotic optimality.

1. Introduction

The choice of the universal Turing machine (UTM) has been a big open question in algorithmic information theory for a long time. While attempts have been made (Müller, 2010) no answer is in sight. The *Kolmogorov complexity* of a string, the length of the shortest program that prints this string, depends on this choice. However, there are *invariance theorems* (Li and Vitányi, 2008, Thm. 2.1.1 & Thm. 3.1.1) which state that changing the UTM changes Kolmogorov complexity only by a constant. When using the *universal prior* M introduced by Solomonoff (1964, 1978) to predict any deterministic computable binary sequence, the number of wrong predictions is bounded by (a multiple of) the Kolmogorov complexity of the sequence (Hutter, 2001). Due to the invariance theorem, changing the UTM changes the number of errors only by a constant. In this sense, compression and prediction work for any choice of UTM.

Hutter (2005) defines the universally intelligent agent AIXI, which is targeted at the *general reinforcement learning problem* (Sutton and Barto, 1998). It extends Solomonoff induction to the interactive setting. AIXI is a Bayesian agent, using a universal prior on the set of all computable environments; actions are taken according to the maximization of expected future discounted rewards. Closely related is the intelligence measure defined by Legg and Hutter (2007), a mathematical performance measure for general reinforcement learning agents: defined as the discounted rewards achieved across all computable environments, weighted by the universal prior.

There are several known positive results about AIXI. It has been proven to be *Pareto optimal* (Hutter, 2002, Thm. 2 & Thm. 6), *balanced Pareto optimal* (Hutter, 2002, Thm. 3), and has

*. This is a short version of a paper appearing at COLT 2015 (Leike and Hutter, 2015b).

maximal Legg-Hutter intelligence. Furthermore, AIXI asymptotically learns to predict the environment perfectly and with a small total number of errors analogously to Solomonoff induction (Hutter, 2005, Thm. 5.36), but only *on policy*: AIXI learns to correctly predict the value (expected future rewards) of its own actions, but generally not the value of counterfactual actions that it does not take.

Orseau (2010, 2013) showed that AIXI does not achieve asymptotic optimality in all computable environments. So instead, we may ask the following weaker questions. Does AIXI succeed in every partially observable Markov decision process (POMDP)/(ergodic) Markov decision process (MDP)/bandit problem/sequence prediction task? We show that without further assumptions on the UTM, we cannot answer any of the preceding questions in the affirmative. More generally, there can be no invariance theorem for AIXI. As a reinforcement learning agent, AIXI has to balance between exploration and exploitation. Acting according to any (universal) prior does not lead to enough exploration, and the bias of AIXI's prior is retained indefinitely. For bad priors this can cause serious malfunctions. However, this problem can be alleviated by adding an extra exploration component to AIXI (Lattimore, 2013, Ch. 5), similar to knowledge-seeking agents (Orseau, 2014; Orseau et al., 2013), or by the use of optimism (Sunehag and Hutter, 2012).

In Section 3 we discuss an example of a universal prior, the *dogmatic prior*, that causes AIXI to misbehave drastically. For any computable policy π the dogmatic prior makes AIXI stick to the policy π as long as expected future rewards do not fall too close to zero. This has profound implications. We show in Section 4 that if we measure Legg-Hutter intelligence with respect to a *different* universal prior, AIXI scores arbitrarily close to the minimal intelligence while any computable policy can score arbitrarily close to the maximal intelligence. This makes the Legg-Hutter intelligence score and thus balanced Pareto optimality relative to the choice of the UTM. Moreover, in Section 5 we show that in the class of all computable environments, *every* policy is Pareto optimal. This undermines all existing optimality results for AIXI. We discuss the implications of these results for optimality notions of general reinforcement learners in Section 6.

2. Preliminaries and Notation

The set $\mathcal{X}^* := \bigcup_{n=0}^{\infty} \mathcal{X}^n$ is the set of all finite strings over the alphabet $\mathcal{X}$, the set $\mathcal{X}^\infty$ is the set of all infinite strings over the alphabet $\mathcal{X}$, and the set $\mathcal{X}^\sharp := \mathcal{X}^* \cup \mathcal{X}^\infty$ is their union. The empty string is denoted by ϵ, not to be confused with the small positive real number ε. Given a string $x \in \mathcal{X}^\sharp$, we denote its length by $|x|$. For a (finite or infinite) string x of length $\geq k$, we denote with $x_{1:k}$ the first k characters of x, and with $x_{<k}$ the first $k-1$ characters of x. The notation $x_{1:\infty}$ stresses that x is an infinite string. We write $x \sqsubseteq y$ iff x is a prefix of y, i.e., $x = y_{1:|x|}$.

In reinforcement learning, the agent interacts with an environment in cycles: at time step t the agent chooses an *action* $a_t \in \mathcal{A}$ and receives a *percept* $e_t = (o_t, r_t) \in \mathcal{E}$ consisting of an *observation* $o_t \in \mathcal{O}$ and a real-valued *reward* $r_t \in \mathbb{R}$; the cycle then repeats for $t+1$. A *history* is an element of $(\mathcal{A} \times \mathcal{E})^*$. We use $æ \in \mathcal{A} \times \mathcal{E}$ to denote one interaction cycle, and $æ_{<t}$ to denote a history of length $t-1$. The goal in reinforcement learning is to maximize total discounted rewards. A *policy* is a function $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$ mapping each history to the action taken after seeing this history. A history $æ_{<t}$ is *consistent with policy* π iff $\pi(æ_{<k}) = a_k$ for all $k < t$.

A function $f : \mathcal{X}^* \to \mathbb{R}$ is *lower semicomputable* iff the set $\{(x, q) \in \mathcal{X}^* \times \mathbb{Q} \mid f(x) > q\}$ is recursively enumerable. A *conditional semimeasure* ν is a probability measure over finite and infinite strings of percepts given actions as input where $\nu(e_{<t} \parallel a_{1:\infty})$ denotes the probability of

receiving percepts $e_{<t}$ when taking actions $a_{1:\infty}$. Formally, ν maps $\mathcal{A}^\infty$ to a probability distribution over $\mathcal{E}^\sharp$.

The conditional semimeasure ν is *chronological* iff the first $t-1$ percepts are independent of future actions a_k for $k \geq t$, i.e., $\nu(e_{<t} \parallel a_{1:k}) = \nu(e_{<t} \parallel a_{<t})$. Despite their name, conditional semimeasures do not denote a conditional probability; ν is not a joint probability distribution over actions and percepts. We model environments as lower semicomputable chronological conditional semimeasures (LSCCCS) (Hutter, 2005, Sec. 5.1.1); the class of all such environments is denoted as $\mathcal{M}_{\text{LSC}}^{\text{CCS}}$. We also use the larger set of all chronological conditional semimeasures $\mathcal{M}^{\text{CCS}}$.

A *universal prior* is a function $w : \mathcal{M}_{\text{LSC}}^{\text{CCS}} \to [0,1]$ such that $w_\nu := w(\nu) > 0$ for all $\nu \in \mathcal{M}_{\text{LSC}}^{\text{CCS}}$ and $\sum_{\nu \in \mathcal{M}_{\text{LSC}}^{\text{CCS}}} w_\nu \leq 1$. A universal prior w gives rise to a *universal mixture*,

$$\xi(e_{<t} \parallel a_{<t}) \;:=\; \sum_{\nu \in \mathcal{M}_{\text{LSC}}^{\text{CCS}}} w_\nu \nu(e_{<t} \parallel a_{<t}). \tag{1}$$

If the universal prior is lower semicomputable, then the universal mixture ξ is an LSCCCS, i.e., $\xi \in \mathcal{M}_{\text{LSC}}^{\text{CCS}}$. From a given universal monotone Turing machine U (Li and Vitányi, 2008, Sec. 4.5.2) we can get a universal mixture ξ in two ways. First, we can use (1) with the prior given by $w_\nu := 2^{-K(\nu)}$, where $K(\nu)$ is the *Kolmogorov complexity* of ν's index in the enumeration of all LSCCCSs (Li and Vitányi, 2008, Eq. 4.11). Second, we can define it as the probability that the universal monotone Turing machine U generates $e_{<t}$ when fed with $a_{<t}$ and uniformly random bits:

$$\xi(e_{<t} \parallel a_{<t}) \;:=\; \sum_{p:\, e_{<t} \sqsubseteq U(p, a_{<t})} 2^{-|p|} \tag{2}$$

Both definitions are equivalent, but not necessarily equal (Wood et al., 2011, Lem. 10 & Lem. 13).

We fix a *discount function* $\gamma : \mathbb{N} \to \mathbb{R}$ with $\gamma_t \geq 0$ and $\sum_{t=1}^\infty \gamma_t < \infty$. The *discount normalization factor* is defined as $\Gamma_t := \sum_{i=t}^\infty \gamma_i$. There is no requirement that $\gamma_t > 0$ or $\Gamma_t > 0$. If $m := \min\{t \mid \Gamma_{t+1} = 0\}$ exists, we say the agent has a *finite lifetime* m and does not care what happens afterwards.

Definition 1 (Value Function) *The* value *of a policy π in an environment ν given history $æ_{<t}$ is defined as $V_\nu^\pi(æ_{<t}) := V_\nu^\pi(æ_{<t}\pi(æ_{<t}))$ and*

$$V_\nu^\pi(æ_{<t}a_t) \;:=\; \frac{1}{\Gamma_t} \sum_{e_t \in \mathcal{E}} \big(\gamma_t r_t + \Gamma_{t+1} V_\nu^\pi(æ_{1:t})\big) \nu(e_{1:t} \mid e_{<t} \parallel a_{1:t})$$

if $\Gamma_t > 0$ and $V_\nu^\pi(æ_{<t}) := 0$ if $\Gamma_t = 0$. The optimal value *is defined as $V_\nu^*(h) := \sup_\pi V_\nu^\pi(h)$.*

Definition 2 (Optimal Policy (Hutter, 2005, Def. 5.19 & 5.30)) *A policy π is* optimal in environment ν (ν-optimal) *iff for all histories the policy π attains the optimal value: $V_\nu^\pi(h) = V_\nu^*(h)$ for all $h \in (\mathcal{A} \times \mathcal{E})^*$. The action $\pi(h)$ is an* optimal action *iff $\pi(h) = \pi_\nu^*(h)$ for some ν-optimal policy π_ν^*.*

Formally, AIXI is defined as a policy π_ξ^* that is optimal in the universal mixture ξ. Since there can be more than one ξ-optimal policy, this definition is not unique. If there two optimal actions $\alpha \neq \beta \in \mathcal{A}$, we call it an *argmax tie*. Which action we take in case of a tie (how we break the tie) is irrelevant and can be arbitrary. We assume that the discount function is summable, rewards are bounded between 0 and 1, and the set of actions $\mathcal{A}$ and the set of percepts $\mathcal{E}$ are both finite. Then an optimal policy exists for every environment $\nu \in \mathcal{M}_{\text{LSC}}^{\text{CCS}}$ (Lattimore and Hutter, 2014, Thm. 10), in particular for any universal mixture ξ.

3. The Dogmatic Prior

In this section we define a universal prior that assigns very high probability of going to hell (reward 0 forever) if we deviate from a given computable policy π. For a Bayesian agent like AIXI, it is thus only worth deviating from the policy π if the agent thinks that the prospects of following π are very poor already. We call this prior the *dogmatic prior*, because the fear of going to hell makes AIXI conform to any arbitrary 'dogmatic ideology' π. AIXI will only break out if it expects π to give very low future payoff; in that case the agent does not have much to lose.

Theorem 3 (Dogmatic Prior) *Let π be any computable policy, let ξ be any universal mixture, and let $\varepsilon > 0$. There is a universal mixture ξ' such that for any history h consistent with π and $V^\pi_\xi(h) > \varepsilon$, the action $\pi(h)$ is the unique ξ'-optimal action.*

The proof proceeds by constructing a universal mixture that assigns disproportionally high probability to an environment ν that sends any policy deviating from π to hell. Importantly, the environment ν produces observations according to the universal mixture ξ. Therefore ν is indistinguishable from ξ on the policy π, so the posterior belief in ν is equal to the prior belief in ν.

Corollary 4 (AIXI Emulating Computable Policies) *Let $\varepsilon > 0$ and let π be any computable policy. There is a universal mixture ξ' such that for any ξ'-optimal policy $\pi^*_{\xi'}$ and for any (not necessarily computable) environment $\nu \in \mathcal{M}^{\mathrm{CCS}}$,*

$$\left| V_\nu^{\pi^*_{\xi'}}(\epsilon) - V_\nu^\pi(\epsilon) \right| \;<\; \varepsilon.$$

Corollary 5 (With Finite Lifetime Every Policy is an AIXI) *If $\Gamma_{m+1} = 0$ for some $m \in \mathbb{N}$, then for any policy π there is a universal mixture ξ' such that $\pi(h)$ is the only ξ'-optimal action for all histories h consistent with π and $|h| \leq m$.*

4. Consequences for Legg-Hutter Intelligence

The aim of the Legg-Hutter intelligence measure is to formalize the intuitive notion of intelligence mathematically. If we take intelligence to mean *an agent's ability to achieve goals in a wide range of environments* (Legg and Hutter, 2007), and we weigh environments according to the universal prior, then the intelligence of a policy π corresponds to the value that π achieves in the corresponding universal mixture. We use the results form the previous section to illustrate some problems with this intelligence measure in the absence of a *natural* UTM.

Definition 6 (Legg-Hutter Intelligence (Legg and Hutter, 2007)) *The* intelligence[1] *of a policy π is defined as*

$$\Upsilon_\xi(\pi) \;:=\; \sum_{\nu \in \mathcal{M}^{\mathrm{CCS}}_{\mathrm{LSC}}} w_\nu V^\pi_\nu(\epsilon) \;=\; V^\pi_\xi(\epsilon).$$

1. Legg and Hutter (2007) consider a subclass of $\mathcal{M}^{\mathrm{CCS}}_{\mathrm{LSC}}$, the class of computable *measures*, and do not use discounting explicitly.

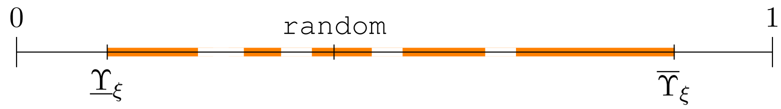

Figure 1: The Legg-Hutter intelligence measure assigns values within the closed interval $[\underline{\Upsilon}_\xi, \overline{\Upsilon}_\xi]$; the assigned values are depicted in orange. By Theorem 7, computable policies are dense in this orange set.

Typically, the index ξ is omitted when writing Υ. However, in this paper we consider the intelligence measure with respect to different universal mixtures, therefore we make this dependency explicit.

Because the value function is scaled to be in the interval $[0, 1]$, intelligence is a real number between 0 and 1. Legg-Hutter intelligence is linked to *balanced Pareto optimality*: a policy is said to be *balanced Pareto optimal* iff it scores the highest intelligence score:

$$\overline{\Upsilon}_\xi \;:=\; \sup_\pi \Upsilon_\xi(\pi) \;=\; \Upsilon_\xi(\pi^*_\xi).$$

AIXI is balanced Pareto optimal (Hutter, 2005, Thm. 5.24). It is just as hard to score very high on the Legg-Hutter intelligence measure as it is to score very low: we can always turn a reward minimizer into a reward maximizer by inverting the rewards $r'_t := 1 - r_t$. Hence the lowest possible intelligence score is achieved by AIXI's twin sister, a ξ-expected reward minimizer:

$$\underline{\Upsilon}_\xi \;:=\; \inf_\pi \Upsilon_\xi(\pi).$$

The heaven environment (reward 1 forever) and the hell environment (reward 0 forever) are computable and thus in the environment class $\mathcal{M}_{\text{LSC}}^{\text{CCS}}$; therefore it is impossible to get a reward 0 or reward 1 in every environment. Consequently, for all policies π,

$$0 \;<\; \underline{\Upsilon}_\xi \;\leq\; \Upsilon_\xi(\pi) \;\leq\; \overline{\Upsilon}_\xi \;<\; 1. \tag{3}$$

See Figure 1. It is natural to fix the policy `random` that takes actions uniformly at random to have an intelligence score of $1/2$ by choosing a 'symmetric' universal prior (Legg and Veness, 2013).

AIXI is not computable (Leike and Hutter, 2015a, Thm. 14), hence there is no computable policy π such that $\Upsilon_\xi(\pi) = \underline{\Upsilon}_\xi$ or $\Upsilon_\xi(\pi) = \overline{\Upsilon}_\xi$ for any universal mixture ξ. But the next theorem tells us that computable policies can come arbitrarily close. This is no surprise: we can do well on a Legg-Hutter intelligence test simply by memorizing what AIXI would do for the first k steps; as long as k is chosen large enough such that discounting makes the remaining rewards contribute very little to the value function.

Theorem 7 (Computable Policies are Dense) $\{\Upsilon_\xi(\pi) \mid \pi$ *is a computable policy*$\}$ *is dense in the set of intelligence scores* $\{\Upsilon_\xi(\pi) \mid \pi$ *is a policy*$\}$.

Remark 8 (Intelligence is not Dense in $[\underline{\Upsilon}_\xi, \overline{\Upsilon}_\xi]$) *The intelligence values of policies are generally not dense in the interval* $[\underline{\Upsilon}_\xi, \overline{\Upsilon}_\xi]$.

Legg-Hutter intelligence is measured with respect to a fixed UTM. AIXI is the most intelligent policy *if it uses the same UTM*. But if we build AIXI with a dogmatic prior, its intelligence score can be arbitrary close to the minimum intelligence score $\underline{\Upsilon}_\xi$.

Corollary 9 (Some AIXIs are Stupid) *For any universal mixture ξ and every $\varepsilon > 0$, there is a universal mixture ξ' such that $\Upsilon_\xi(\pi^*_{\xi'}) < \underline{\Upsilon}_\xi + \varepsilon$.*

We get the same result if we fix AIXI, but rig the intelligence measure.

Corollary 10 (AIXI is Stupid for Some Υ) *For any ξ-optimal policy π^*_ξ and for every $\varepsilon > 0$ there is a universal mixture ξ' such that $\Upsilon_{\xi'}(\pi^*_\xi) \leq \varepsilon$ and $\overline{\Upsilon}_{\xi'} \geq 1 - \varepsilon$.*

On the other hand, we can make any computable policy smart if we choose the right universal mixture. In particular, we get that there is a universal mixture such that 'do nothing' is the most intelligent policy save for some ε!

Corollary 11 (Computable Policies can be Smart) *For any computable policy π and any $\varepsilon > 0$ there is a universal mixture ξ' such that $\Upsilon_{\xi'}(\pi) > \overline{\Upsilon}_{\xi'} - \varepsilon$.*

5. Pareto Optimality

In Section 3 we have seen an example for a bad universal prior. But we know that for any universal prior, AIXI is *Pareto optimal* (Hutter, 2002). Here we show that Pareto optimality is not a useful criterion for optimality since for any environment class containing $\mathcal{M}^{\mathrm{CCS}}_{\mathrm{LSC}}$, all policies are Pareto optimal.

Definition 12 (Pareto Optimality (Hutter, 2005, Def. 5.22)) *Let $\mathcal{M}$ be a set of environments. A policy π is* Pareto optimal in the set of environments $\mathcal{M}$ *iff there is no policy $\tilde{\pi}$ such that $V^{\tilde{\pi}}_\nu(\epsilon) \geq V^\pi_\nu(\epsilon)$ for all $\nu \in \mathcal{M}$ and $V^{\tilde{\pi}}_\rho(\epsilon) > V^\pi_\rho(\epsilon)$ for at least one $\rho \in \mathcal{M}$.*

Theorem 13 (AIXI is Pareto Optimal (Hutter, 2005, Thm. 5.32)) *A ξ-optimal policy is Pareto optimal in $\mathcal{M}^{\mathrm{CCS}}_{\mathrm{LSC}}$.*

Theorem 14 (Pareto Optimality is Trivial) *Every policy is Pareto optimal in any $\mathcal{M} \supseteq \mathcal{M}^{\mathrm{CCS}}_{\mathrm{LSC}}$.*

The proof proceeds as follows: for a given policy π, we construct a set of 'buddy environments' that reward π and punish other policies. Together they can defend against any policy $\tilde{\pi}$ that tries to take the crown of Pareto optimality from π.

6. Discussion

Bayesian reinforcement learning agents make the trade-off between exploration and exploitation in the Bayes-optimal way. The amount of exploration this incurs varies wildly: the dogmatic prior defined in Section 3 can prevent a Bayesian agent *from taking a single exploratory action*; exploration is restricted to cases where the expected future payoff falls below some prespecified $\varepsilon > 0$.

In the introduction we raised the question of whether AIXI succeeds in various subclasses of all computable environments. Interesting subclasses include sequence prediction tasks, (ergodic) (PO)MDPs, bandits, etc. Using a dogmatic prior (Theorem 3), we can make AIXI follow any computable policy as long as that policy produces rewards that are bounded away from zero.

Name	Issue/Comment
μ-optimal policy	requires to know the true environment μ in advance
Pareto optimality	trivial (Theorem 14)
Balanced Pareto optimality	dependent on UTM (Corollary 9 and Corollary 10)
Self-optimizing	does not apply to $\mathcal{M}_{\mathrm{LSC}}^{\mathrm{CCS}}$
Strong asymptotic optimality	impossible (Lattimore and Hutter, 2011, Thm. 8)
Weak asymptotic optimality	BayesExp (Lattimore, 2013, Ch. 5), but not AIXI (Orseau, 2010)

Table 1: Proposed notions of optimality (Hutter, 2002; Orseau, 2010; Lattimore and Hutter, 2011) and their issues. Weak asymptotic optimality stands out to be the only possible nontrivial optimality notion.

- In a sequence prediction task that gives a reward of 1 for every correctly predicted bit and 0 otherwise, a policy π that correctly predicts every third bit will receive an average reward of $1/3$. With a π-dogmatic prior, AIXI thus only predicts a third of the bits correctly, and hence is outperformed by a uniformly random predictor.

  However, if we have a constant horizon of 1, AIXI *does* succeed in sequence prediction (Hutter, 2005, Sec. 6.2.2). If the horizon is this short, the agent is so hedonistic that no threat of hell can deter it.

- In a (partially observable) Markov decision process, a dogmatic prior can make AIXI get stuck in any loop that provides nonzero expected rewards.

- In a bandit problem, a dogmatic prior can make AIXI get stuck on any arm which provides nonzero expected rewards.

These results apply not only to AIXI, but generally to Bayesian reinforcement learning agents. Any Bayesian mixture over reactive environments is susceptible to dogmatic priors if we allow an arbitrary reweighing of the prior. A notable exception is the class of all ergodic MDPs with an unbounded effective horizon; here the Bayes-optimal policy is *strongly asymptotically optimal* (Hutter, 2005, Thm. 5.38): $V_\mu^\pi(æ_{<t}) - V_\mu^*(æ_{<t}) \to 0$ as $t \to \infty$ for all histories $æ_{<t}$.

Moreover, Bayesian agents might still perform well at learning: AIXI's posterior belief about the value of its own policy π_ξ^* converges to the true value while following that policy (Hutter, 2005, Thm. 5.36):

$$V_\xi^{\pi_\xi^*}(æ_{<t}) - V_\mu^{\pi_\xi^*}(æ_{<t}) \to 0 \text{ as } t \to \infty \text{ } \mu\text{-almost surely.}$$

This means that the agent learns to predict those parts of the environment that it sees. But if it does not explore enough, then it will not learn other parts of the environment that are potentially more rewarding.

Theorem 14 proves that Pareto optimality is trivial in the class of all computable environments; Corollary 9 and Corollary 10 show that maximal Legg-Hutter intelligence (balanced Pareto optimality) is highly subjective, because it depends on the choice of the UTM: AIXI is not balanced Pareto optimal with respect to all universal mixtures. Moreover, according to Corollary 11, any computable policy is nearly balanced Pareto optimal, save some $\varepsilon > 0$. The self-optimizing theorem (Hutter, 2002, Thm. 4 & Thm. 7) is not applicable to the class of all computable environments

$\mathcal{M}_{\text{LSC}}^{\text{CCS}}$ that we consider here, since this class does not allow for self-optimizing policies. Therefore no nontrivial and non-subjective optimality results for AIXI remain (see Table 1). We have to regard AIXI as a *relative* theory of intelligence, dependent on the choice of the UTM (Sunehag and Hutter, 2014).

The underlying problem is that a discounting Bayesian agent such as AIXI does not have enough time to explore sufficiently; exploitation has to start as soon as possible. In the beginning the agent does not know enough about its environment and therefore relies heavily on its prior. Lack of exploration then retains the prior's biases. This fundamental problem can be alleviated by adding an extra exploration component. Lattimore (2013) defines BayesExp, a *weakly asymptotically optimal policy* π that converges (for all universal priors) to the optimal value in Cesàro mean:

$$\frac{1}{t}\sum_{k=1}^{t}\left(V_\nu^*(æ_{<k}) - V_\nu^\pi(æ_{<k})\right) \xrightarrow{t\to\infty} 0 \quad \nu\text{-almost surely for all } \nu \in \mathcal{M}_{\text{LSC}}^{\text{CCS}}.$$

But it is not clear that weak asymptotic optimality is a good optimality criterion. For example, weak asymptotic optimality can be achieved by navigating into traps (parts of the environment with a simple optimal policy but possibly very low rewards that cannot be escaped). Furthermore, to be weakly asymptotically optimal requires an excessive amount of exploration: BayesExp needs to take exploratory actions that it itself knows to very likely be extremely costly or dangerous. This leaves us with the following open question: *What are good optimality criteria for generally intelligent agents* (Hutter, 2009, Sec. 5)?

References

Marcus Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Sciences*, 62(4):653–667, 2001.

Marcus Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Computational Learning Theory*, pages 364–379. Springer, 2002.

Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

Marcus Hutter. Open problems in universal induction & intelligence. *Algorithms*, 3(2):879–906, 2009.

Tor Lattimore. *Theory of General Reinforcement Learning*. PhD thesis, Australian National University, 2013.

Tor Lattimore and Marcus Hutter. Asymptotically optimal agents. In *Algorithmic Learning Theory*, pages 368–382. Springer, 2011.

Tor Lattimore and Marcus Hutter. General time consistent discounting. *Theoretical Computer Science*, 519:140–154, 2014.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds & Machines*, 17(4):391–444, 2007.

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.

Jan Leike and Marcus Hutter. On the computability of AIXI. In *Uncertainty in Artificial Intelligence*, 2015a.

Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. In *Conference on Learning Theory*, 2015b.

Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer, 3rd edition, 2008.

Markus Müller. Stationary algorithmic probability. *Theoretical Computer Science*, 411(1):113–130, 2010.

Laurent Orseau. Optimality issues of universal greedy agents with static priors. In *Algorithmic Learning Theory*, pages 345–359. Springer, 2010.

Laurent Orseau. Asymptotic non-learnability of universal agents with computable horizon functions. *Theoretical Computer Science*, 473:149–156, 2013.

Laurent Orseau. Universal knowledge-seeking agents. *Theoretical Computer Science*, 519:127–139, 2014.

Laurent Orseau, Tor Lattimore, and Marcus Hutter. Universal knowledge-seeking agents for stochastic environments. In *Algorithmic Learning Theory*, pages 158–172. Springer, 2013.

Ray Solomonoff. A formal theory of inductive inference. Parts 1 and 2. *Information and Control*, 7 (1):1–22 and 224–254, 1964.

Ray Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, 24(4):422–432, 1978.

Peter Sunehag and Marcus Hutter. Optimistic agents are asymptotically optimal. In *Australasian Joint Conference on Artificial Intelligence*, pages 15–26. Springer, 2012.

Peter Sunehag and Marcus Hutter. Intelligence as inference or forcing Occam on the world. In *Artificial General Intelligence*, pages 186–195. Springer, 2014.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Ian Wood, Peter Sunehag, and Marcus Hutter. (Non-)equivalence of universal priors. In *Solomonoff 85th Memorial Conference*, pages 417–425. Springer, 2011.