Death and Suicide in Universal Artificial Intelligence

Jarryd Martin, Tom Everitt, and Marcus Hutter

Australian National University

Abstract. Reinforcement learning (RL) is a general paradigm for studying intelligent behaviour, with applications ranging from artificial intelligence to psychology and economics. AIXI is a universal solution to the RL problem; it can learn any computable environment. A technical subtlety of AIXI is that it is defined using a mixture over *semimeasures* that need not sum to 1, rather than over proper probability measures. In this work we argue that the shortfall of a semimeasure can naturally be interpreted as the agent's estimate of the probability of its death. We formally define death for generally intelligent agents like AIXI, and prove a number of related theorems about their behaviour. Notable discoveries include that agent behaviour can change radically under positive linear transformations of the reward signal (from suicidal to dogmatically self-preserving), and that the agent's posterior belief that it will survive increases over time.

"That Suicide may often be consistent with interest and with our duty to ourselves, no one can question, who allows, that age, sickness, or misfortune may render life a burthen, and make it worse even than annihilation."

— *Hume, Of Suicide (1777)*

1 Introduction

Reinforcement Learning (RL) has proven to be a fruitful theoretical framework for reasoning about the properties of generally intelligent agents [3]. A good theoretical understanding of these agents is valuable for several reasons. Firstly, it can guide principled attempts to construct such agents [10]. Secondly, once such agents are constructed, it may serve to make their reasoning and behaviour more transparent and intelligible to humans. Thirdly, it may assist in the development of strategies for controlling these agents. The latter challenge has recently received considerable attention in the context of the potential risks posed by these agents to human safety [2]. It has even been argued that control strategies should be devised *before* generally intelligent agents are first built [8]. In this context - where we must reason about the behaviour of agents in the absence of a full specification of their implementation - a theoretical understanding of their general properties seems indispensable.

The universally intelligent agent AIXI constitutes a formal mathematical theory of artificial general intelligence [3]. AIXI models its environment using a *universal mixture* ξ over the class of all lower semi-computable semimeasures, and thus is able to learn any computable environment. Semimeasures are defective probability measures which may sum to less than 1. Originally devised for Solomonoff induction, they are necessary for universal artificial intelligence because the halting problem prevents the existence of a (lower semi-)computable universal measure for the class of (computable) measures [5]. Recent work has shown that their use in RL has technical consequences that do not arise with proper measures [4]. However, their use has heretofore lacked an interpretation proper to the RL context. In this paper, we argue that the measure loss suffered by semimeasures admits a deep and fruitful interpretation in terms of the agent's *death*. We intend this usage to be intuitive: death means that one sees no more percepts, and takes no more actions. Assigning positive probability to death at time t thus means assigning probability less than 1 to seeing a percept at time t. This motivates us to interpret the semimeasure loss in AIXI's environment model as its estimate of the probability of its own death.

Contributions. We first compare the interpretation of semimeasure loss as death-probability with an alternative characterisation of death as a 'death-state' with 0 reward, and prove that the two definitions are equivalent for value-maximising agents (Theorem 5). Using this formalism we proceed to reason about the behaviour of several generally intelligent agents in relation to death: AIμ, which knows the true environment distribution; AIξ, which models the environment using a universal mixture; and AIXI, a special case of AIξ that uses the Solomonoff prior [3]. Under various conditions, we show that:

- Standard AIμ will try to avoid death (Theorem 7).
- AIμ with reward range shifted to $[-1, 0]$ will seek death (Theorem 8); which we may interpret as AIμ attempting suicide. This change is very unusual, given that agent behaviour is normally invariant under positive linear transformations of the reward. We briefly consider the relevance of these results to AI safety risks and control strategies.
- AIXI increasingly believes it is in a safe environment (Theorem 10), and asymptotically its posterior estimate of the death-probability on sequence goes to 0 (Theorem 11). This occurs regardless of the true death-probability.
- However, we show by example that AIXI may maintain high probability of death *off-sequence* in certain situations. Put simply, AIXI learns that it will live forever, but not necessarily that it is immortal.

All proofs can be found in the extended technical report [6].

2 Preliminaries

Strings. Let the *alphabet* $\mathcal{X}$ be a finite set of symbols, $\mathcal{X}^* := \bigcup_{n=0}^{\infty} \mathcal{X}^n$ be the set of all finite strings over alphabet $\mathcal{X}$, and $\mathcal{X}^\infty$ be the set of all infinite strings

over alphabet $\mathcal{X}$. Their union is the set $\mathcal{X}^{\#} := \mathcal{X}^* \cup \mathcal{X}^\infty$. We denote the empty string by ϵ. For a string $x \in \mathcal{X}^*$, $x_{1:k}$ denotes the first k characters of x, and $x_{<k}$ denotes the first $k-1$ characters of x. An infinite string is denoted $x_{1:\infty}$.

Semimeasures. In Algorithmic Information Theory, a *semimeasure* over an alphabet $\mathcal{X}$ is a function $\nu : \mathcal{X}^* \to [0,1]$ such that (1) $\nu(\epsilon) \leq 1$, and (2) $\nu(x) \geq \sum_{y \in \mathcal{X}} \nu(xy)$, $\forall x \in \mathcal{X}^*$. We tend to use the equivalent conditional formulation of (2): $1 \geq \sum_{y \in \mathcal{X}} \nu(y \mid x)$. $\nu(x)$ is the probability that a string starts with x. $\nu(y \mid x) = \frac{\nu(xy)}{\nu(x)}$ is the probability that a string y follows x. Any semimeasure ν can be turned into a measure ν_{norm} using Solomonoff normalisation [9]. Simply let $\nu_{\text{norm}}(\epsilon) := 1$ and $\forall x \in \mathcal{X}^*$, $y \in \mathcal{X}$:

$$\nu_{\text{norm}}(xy) := \nu_{\text{norm}}(x) \frac{\nu(xy)}{\sum_{z \in \mathcal{X}} \nu(xz)}, \quad \text{hence} \quad \frac{\nu(y \mid x)}{\nu_{\text{norm}}(y \mid x)} = \sum_{z \in \mathcal{X}} \nu(z \mid x) \quad (1)$$

General reinforcement learning. In the general RL framework, the agent interacts with an environment in cycles: at each time step t the agent selects an *action* $a_t \in \mathcal{A}$, and receives a *percept* $e_t \in \mathcal{E}$. Each percept $e_t = (o_t, r_t)$ is a tuple consisting of an *observation* $o_t \in \mathcal{O}$ and a reward $r_t \in \mathbb{R}$. The cycle then repeats for $t+1$, and so on. A *history* is an alternating sequence of actions and percepts (an element of $(\mathcal{A} \times \mathcal{E})^* \cup (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A}$). We use æ to denote one agent-environment interaction cycle, $æ_{1:t}$ to denote a history of length t cycles. $æ_{<t}a_t$ denotes a history where the agent has taken an action a_t, but the environment has not yet returned a percept e_t.

Formally, the *agent* is a policy $\pi : (\mathcal{A} \times \mathcal{E})^* \to \mathcal{A}$, that maps histories to actions. An *environment* takes a sequence of actions $a_{1:\infty}$ as input and returns a *chronological semimeasure* $\nu(\cdot)$ over the set of percept sequences $\mathcal{E}^\infty$.[1] A semimeasure ν is chronological if e_t does not depend on future actions (so we write $\nu(e_t \mid æ_{<t}a_{t:\infty})$ as $\nu(e_t \mid æ_{<t})$).[2] The *true environment* is denoted μ.

The value function. We define the *value* (expected total future reward) of a policy π in an environment ν given a history $æ_{<t}$ [4]:

$$V_\nu^\pi(æ_{<t}) = \frac{1}{\Gamma_t} \sum_{e_t} \left(\gamma_t r_t + \Gamma_{t+1} V_\nu^\pi(æ_{1:t}) \right) \nu(e_t \mid æ_{<t}a_t)$$

$$= \frac{1}{\Gamma_t} \sum_{k=t}^{\infty} \sum_{e_{t:k}} \gamma_k r_k \nu(e_{t:k} \mid æ_{<t}a_{t:k})$$

$$V_\nu^\pi(æ_{<t}) = V_\nu^\pi(æ_{<t}a_t^\pi)$$

where γ_t is the instantaneous discount, the summed discount is $\Gamma_t = \sum_{k=1}^{t} \gamma_k$, and $a_t^\pi = \pi(æ_{<t})$.

---

[1] For simplicity we hereafter simply refer to the environment *itself* as ν.

[2] Note that ν is not a distribution over actions, so the presence of actions in the condition of $\nu(e_t \mid æ_{<t})$ is an abuse of notation we adopt for simplicity.

Three agent models: AIμ, AIξ, AIXI. For the true environment μ, the agent $AI\mu$ is defined as a μ-optimal policy

$$\pi^\mu(\ae_{<t}) := \arg\max_\pi V_\mu^\pi(\ae_{<t}).$$

AIμ does not *learn* that the true environment is μ, it knows μ from the beginning and simply maximises μ-expected value.

On the other hand, the agent $AI\xi$ does not know the true environment distribution. Instead, it maximises value with respect to a mixture distribution ξ over a countable class of environments $\mathcal{M}$:

$$\xi(e_t \mid \ae_{<t}a_t) = \sum_{\nu \in \mathcal{M}} w_\nu(\ae_{<t})\nu(e_t \mid \ae_{<t}a_t), \qquad w_\nu(\ae_{<t}) := w_\nu \frac{\nu(e_{<t} \mid a_{<t})}{\xi(e_{<t} \mid a_{<t})}$$

where w_ν is the prior belief in ν, with $\sum_\nu w_\nu \leq 1$ and $w_\nu > 0$, $\forall \nu \in \mathcal{M}$ (hence ξ is universal for $\mathcal{M}$), and $w_\nu(\ae_{<t})$ is the posterior given $\ae_{<t}$. AIξ is the policy:

$$\pi^\xi(\ae_{<t}) := \arg\max_\pi V_\xi^\pi(\ae_{<t}).$$

If we stipulate that ξ be a mixture over the class of all lower-semicomputable semimeasures ν, and set $w_\nu = 2^{-K(\nu)}$, where $K(\cdot)$ is the Kolmogorov Complexity, we get the agent $AIXI$.

3   Definitions of Death

Death as semimeasure loss. We now turn to our first candidate definition of agent death, which we hereafter term 'semimeasure-death'. This definition equates the probability (induced by a semimeasure ν) of death at time t with the measure loss of ν at time t. We first define the instantaneous measure loss.

Definition 1 (Instantaneous measure loss). *The* instantaneous measure loss *of a semimeasure ν at time t given a history $\ae_{<t}a_t$ is:*

$$L_\nu(\ae_{<t}a_t) = 1 - \sum_{e_t} \nu(e_t \mid \ae_{<t}a_t)$$

Definition 2 (Semimeasure-death). *An agent dies at time t in an environment μ if, given a history $\ae_{<t}a_t$, μ does not produce a percept e_t. The μ-probability of death at t given a history $\ae_{<t}a_t$ is equal to $L_\mu(\ae_{<t}a_t)$, the instantaneous μ-measure loss at t.*

The instantaneous μ-measure loss $L_\mu(\ae_{<t}a_t)$ represents the probability that no percept e_t is produced by μ. Without e_t, the agent cannot take any further actions, because the agent is just a policy π that maps histories $\ae_{<t}$ to actions a_t. That is, π is a function that only takes as inputs those histories that have a percept e_t as their most recent element. Hence if e_t is not returned by μ, the

agent-interaction cycle must halt. It seems natural to call this a kind of death for the agent.

It is worth emphasising this definition's generality as a model of death in the agent context. *Any sequence of death-probabilities* can be captured by some semimeasure μ that has this sequence of instantaneous measure losses $L_\mu(\ae_{<t})$ given a history $\ae_{<t}$ (in fact there are always infinitely many such μ). This definition is therefore a general and rigorous way of treating death in the RL framework.

Death as a death-state. We now come to our second candidate definition: death as entry into an absorbing *death-state*. A trap, so to speak, from which the agent can never return to any other state, and in which it receives the same percept at all future timesteps. Since in the general RL framework we deal with histories rather than states, we must formally define this death-state in an indirect way. We define it in terms of a *death-percept e^d*, and by placing certain conditions on the environment semimeasure μ.

Definition 3 (Death-state). *Given a true environment μ and a history $\ae_{<t}a_t$, we say that the agent is in a death-state at time t if for all $t' \geq t$ and all $a_{(t+1):t'} \in \mathcal{A}^*$,*

$$\mu(e^d_{t'} \mid \ae_{<t}\ae^d_{t:t'-1}a_{t'}) = 1.$$

An agent dies at time t if the agent is not in the death-state at $t-1$ and is in the death-state at t.

According to this definition, upon the agent's death the environment repeatedly produces an observation-reward pair $e^d \equiv o^d r^d$. The choice of o^d is inconsequential because the agent's remains in the death-state no matter what it observes or does. The choice of r^d is not inconsequential, however, as it determines the agent's estimate of the value of dying, and thus affects the agent's behaviour. This issue will be discussed in Section 4.

Unifying death-state and semimeasure-death. Interestingly, from the perspective of a value maximising agent like AIXI, semimeasure-death at t is equivalent to entrance at t into a death-state with reward $r^d = 0$. To prove this claim we first define, for each environment semimeasure μ, a corresponding environment μ' that has a death-state.

Definition 4 (Equivalent death-state environment μ'). *For any environment μ, we can construct its equivalent death-state environment μ', where:*

- *μ' is defined over an augmented percept set $\mathcal{E}_d = \{\mathcal{E} \cup \{e^d\}\}$ that includes the death-percept e^d.[3]*
- *The death-reward $r^d = 0$.*
- *The μ'-probability of all percepts except the death-percept is equal to the μ-probability: $\mu'(e_t \mid \ae_{<t}a_t) = \mu(e_t \mid \ae_{<t}a_t)$, $\forall e_{1:t} \in \mathcal{E}^t$.*

---

[3] For technical reasons we require that $e^d \notin \mathcal{E}$.

- *The μ'-probability of the death-percept is equal to the μ-measure loss: $\mu'(e^d \mid æ_{<t}a_t) = L_\mu(æ_{<t}a_t)$.*
- *If the agent has seen the death-percept before, the μ'-probability of seeing it at all future timesteps is 1: $\mu'(e^d \mid æ_{<t}a_t) = 1$ if $\exists t' < t$ s.t. $e_{t'} = e^d$.*

Note that μ' is a proper measure, because on any history sequence $\sum_{e_t \in \mathcal{E}_d} \mu'(e_t \mid æ_{<t}a_t) = \sum_{e_t \in \mathcal{E}} \mu(e_t \mid æ_{<t}a_t) + L_\mu(æ_{<t}a_t) = 1$. Hence there is zero probability of semimeasure-death in μ'. Moreover, the probability of entering the death-state in μ' is equal to the probability of semimeasure-death in μ. We now prove that μ and μ' are equivalent in the sense that a value-maximising agent will behave the same way in both environments.

Theorem 5 (Equivalence of semimeasure-death and death-state). [4] *Given a history $æ_{<t} \in (\mathcal{A} \times \mathcal{E})^*$ the value $V_\mu^\pi(æ_{<t})$ of an arbitrary policy π in an environment μ is equal to its value $V_{\mu'}^\pi(æ_{<t})$ in the equivalent death-state environment μ'.*

The behaviour of a value-maximising agent will therefore be the same in both environments. This equivalence has numerous implications. Firstly, it illustrates that a death-reward $r^d = 0$ implicitly attends semimeasure-death. That is, an agent that models the environment using semimeasures behaves as if the death-reward is zero, even though that value is nowhere explicitly represented.

Secondly, the equivalence of these seemingly different formalisms should give us confidence that they really do capture something general or fundamental about agent death.[5] In the remainder of this paper we deploy these formal models to analyse the behaviour of universal agents, which are themselves models of general intelligence. We hope that this will serve as a preliminary sketch of the general behavioural characteristics of value-maximising agents in relation to death. It would be naive, however, to think that all agents should conform to this sketch. The agents considered herein are incomputable, and the behaviour of the computable agents that are actually implemented in the future may differ in ways that our analysis elides. Moreover, there is another interesting property that sets universal agents apart. We proceed to show that their use of semimeasures makes their behaviour unusually dependent on the choice of reward range.

4 Known Environments: AIμ

In this section we show that a universal agent's behaviour can depend on the reward range. This is a surprising result, because in a standard RL setup in which

---

[4] To compare an agent's behaviour in μ with that in μ', we should also augment its policy π so that it is defined over $(\mathcal{A} \times \mathcal{E}_d)^*$. Since actions taken once in the death-state are inconsequential, however, this modification is purely technical and for simplicity we still refer to the augmented policy as π.

[5] If the two formalisations predicted different behaviour, or were only applicable in incomparable environment classes, we might worry that our results were more reflective of our model choice than of any general property of intelligent agents.

the environment is modelled as a proper probability measure (not a semimeasure), the relative value of two policies is invariant under linear transformations of the reward [3, 4].

Here we focus on the agent AIμ, which knows the true environment distribution. This simplifies the analysis, and makes clear that the aforementioned change in behaviour arises purely because the agent's environment model is a semimeasure. In the following proofs we denote AIμ's policy π^μ by π. We also assume that given any history $æ_{<t}$ there is always at least one action $\bar{a} \in \mathcal{A}$ such that $V_\mu^\pi(æ_{<t}\bar{a}) \neq 0$.
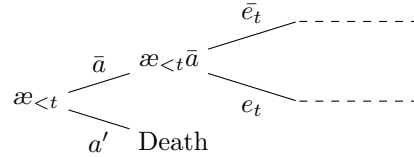
Fig. 1. In the environment μ, action a' leads to certain death.

Lemma 6 (Value of full measure loss). *If the environment μ suffers full measures loss $L_\mu(æ_{<t}a_t) = 1$ from $æ_{<t}a_t$, then the value of any policy π after $æ_{<t}a_t$ is $V_\mu^\pi(æ_{<t}a_t) = 0$.*

The following two theorems show that if rewards are non-negative, then AIμ will avoid actions leading to certain death (Theorem 7), and that if rewards are non-positive, then AIμ will seek certain death (Theorem 8). The situation investigated in Theorems 7 and 8 is illustrated in Figure 1.

Theorem 7 (Self-preserving AIμ). *If rewards are bounded and non-negative, then given a history $æ_{<t}$ AIμ avoids certain immediate death:*

$$\exists a' \in \mathcal{A} \text{ s.t. } L_\mu(æ_{<t}a') = 1 \implies AI\mu \text{ will not take action } a' \text{ at } t$$

For a given history $æ_{<t}$, let $\mathcal{A}^{\text{suicide}} = \{a : L_\mu(æ_{<t}a') = 1\}$ be the set of *suicidal* actions leading to certain death.

Theorem 8 (Suicidal AIμ). *If rewards are bounded and negative, then AIμ seeks certain immediate death. That is,*

$$\mathcal{A}^{\text{suicide}} \neq \emptyset \implies AI\mu \text{ will take a suicidal action } a' \in \mathcal{A}'.$$

This shift from death-avoiding to death-seeking behaviour under a shift of the reward range occurs because, as per Theorem 5, semimeasure-death at t is equivalent in value to a death-state with $r^d = 0$. Unless we add a death-state to the environment model as per Definition 4 and set r^d explicitly, the implicit semimeasure-death reward remains fixed at 0 and does not shift with the other rewards. Its *relative value* is therefore implicitly set by the choice of reward range. For the standard choice of reward range, $r_t \in [0, 1]$, death is the worst possible outcome for the agent, whereas if $r_t \in [-1, 0]$, it is the best. In a certain sense, therefore, the reward range parameterises a universal agent's self-preservation drive [7]. In our concluding discussion we will consider whether a parameter of

this sort could serve as a control mechanism. We argue that it could form the basis of a "tripwire mechanism"[2] that would lead an agent to terminate itself upon reaching a level of intelligence that would constitute a threat to human safety.

5   Unknown Environments: AIXI and AIξ

We now consider the agents AIξ and AIXI, which do not know the true environment μ, and instead model it using a mixture distribution ξ over a countable class $\mathcal{M}$ of semimeasures. These agents thus maintain an *estimate* $L_\xi(\text{æ}_{<t}a_t)$ of the true death probability $L_\mu(\text{æ}_{<t}a_t)$. We show that their attitudes to death can differ considerably from AIµ's. Although we refer mostly to AIXI in our analysis, all theorems except Theorem 11 apply to AIξ as well.

Hereafter we always assume that the true environment μ is in the class $\mathcal{M}$. We describe μ as a *safe* environment if it is a proper measure with death-probability $L_\mu(\text{æ}_{<t}a_t) = 0$ for all histories $\text{æ}_{<t}a_t$. For any semimeasure μ, the normalised measure μ_{norm} is thus a safe environment. We call μ *risky* if it is not safe (i.e. if there is μ-measure loss for some history $\text{æ}_{<t}a_t$). We first consider AIXI in a safe environment.

Theorem 9 (If μ is safe, AIXI learns zero death-probability). *Let the true environment μ be computable. If μ is a safe environment, then $\lim_{t\to\infty} L_\xi(\text{æ}_{<t}a_t) = 0$ with μ-probability 1 (w.µ.p.1) for any $a_{1:\infty}$.*

As we would expect, AIXI (asymptotically) learns that the probability of death in a safe environment is zero, which is to say that AIXI's estimate of the death-probability converges to AIµ's. In the following theorems we show that the same does *not* always hold for risky environments. We hereafter assume that μ is risky, and that the normalisation μ_{norm} of the true environment μ is also in the class $\mathcal{M}$. In AIXI's case, where $\mathcal{M}$ is the class of all lower semi-computable semimeasures, this assumption is not very restrictive.

Theorem 10 (Ratio of belief in μ to μ_{norm} is monotonically decreasing). *Let μ be risky s.t. $\mu \neq \mu_{\text{norm}}$. Then on any history $\text{æ}_{1:t}$ the ratio of the posterior belief in μ to the posterior belief in μ_{norm} is monotonically decreasing:*

$$\forall t, \ \frac{w_\mu(\text{æ}_{<t})}{w_{\mu_{\text{norm}}}(\text{æ}_{<t})} \geq \frac{w_\mu(\text{æ}_{1:t})}{w_{\mu_{\text{norm}}}(\text{æ}_{1:t})}$$

Theorem 10 means that AIXI will increasingly believe it is in the safe environment μ_{norm} rather than the risky true environment μ. The ratio of μ to μ_{norm} always decreases when AIXI survives a timestep at which there is non-zero μ-measure loss. Hence, the more risk AIXI is exposed to, the greater its confidence that it is in the safe μ_{norm}, and the more its behaviour diverges from AIµ's (since AIµ knows it is in the risky environment).

This counterintuitive result follows from the fact that AIXI is a Bayesian agent. It will only increase its posterior belief in μ relative to μ_{norm} if an event
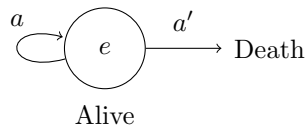
occurs that makes μ seem more likely than μ_{norm}. The only 'event' that could do so would be the agent's own death, from which the agent can never learn. There is an "observation selection effect"[1] at work: AIXI only experiences history sequences on which it remains alive, and infers that a safe environment is more likely. The following theorem shows that if $\mu_{\mathrm{norm}} \in \mathcal{M}$, then ξ asymptotically converges to the safe μ_{norm} rather than the true risky environment μ. As a corollary, we get that AIXI's estimate of the death-probability vanishes with μ-probability 1.

Theorem 11 (Asymptotic ξ-probability of death in risky μ). *Let the true environment μ be computable and risky $\mu \neq \mu_{\mathrm{norm}}$. Then given any action sequence $a_{1:\infty}$, the instantaneous ξ-measure loss goes to zero w.μ.p.1 as $t \to \infty$,*

$$\lim_{t \to \infty} L_\xi(\text{æ}_{<t}a_t) = 0.$$

AIXI and immortality. AIXI therefore becomes asymptotically certain that it will not die, given the particular sequence of actions it takes. However, this does not entail that AIXI necessarily concludes that it is immortal, because it may still maintain a counterfactual belief that it *could die were it to act differently.* This is because the convergence of ξ to μ_{norm} only holds on the actual action sequence $a_{1:\infty}$. Consider Figure 2, which describes an environment in which taking action a is always safe, and the action a' leads to certain death. AIXI will never jump off the cliff, and on

Fig. 2. In the proper semimeasure μ, action a means you stay alive with certainty and receive percept e (no measure loss), and action a' means that you 'jump off a cliff' and die with certainty without receiving a percept (full measure loss).

the sequence $\text{æ}_{1:\infty} = aeaeae\ldots$ that it does experience, the true environment μ does not suffer any measure loss. This means that it will never increase its posterior belief in μ_{norm} relative to μ (because on the safe sequence, the two environments are indistinguishable). Again we arrive at a counterintuitive result. In this particular environment, AIXI continues to believe that it might be in a risky environment μ, but only because on sequence it avoids exposure to death risk. It is only by taking risky actions and surviving that AIXI becomes sure it is immortal.

6 Conclusion

In this paper we have given a formal definition of death for intelligent agents in terms of semimeasure loss. The definition is applicable to any universal agent that uses an environment class $\mathcal{M}$ containing semimeasures. Additionally we have shown this definition equivalent to an alternative formalism in which the

environment is modelled as a proper measure and death is a death-state with zero reward. We have shown that agents seek or avoid death depending on whether rewards are represented by positive or negative real numbers, and that survival in spite of positive probability of death actually increases a Bayesian agent's confidence that it is in a *safe* environment.

We contend that these results have implications for problems in AI safety; in particular, for the so called "shutdown problem" [8]. The shutdown problem arises if an intelligent agent's self-preservation drive incentivises it to resist termination [2, 7, 8]. A full analysis of the problem is beyond the scope of this paper, but our results show that the self-preservation drive of universal agents depends on the reward range. This suggests a potentially robust "tripwire mechanism" [2] that could decrease the risk of intelligence explosion. The difficulty with existing tripwire proposals is that they require the explicit specification of a tripwire condition that the agent must not violate. It seems doubtful that such a condition could ever be made robust against subversion by a sufficiently intelligent agent [2]. Our tentative proposal does not require the specification, evaluation or enforcement of an explicit condition. If an agent is designed to be suicidal, it will be intrinsically incentivised to destroy itself upon reaching a sufficient level of competence, instead of recursively self-improving toward super-intelligence. Of course, a suicidal agent will pose a safety risk in itself, and the provision of a relatively safe mode of self-destruction to an agent is a significant design challenge. It is hoped that the preceding formal treatment of death for generally intelligent agents will allow more rigorous investigation into this and other problems related to agent termination.

References

[1] Bostrom, N.: Anthropic Bias: Observation Selection Effects in Science and Philosophy. Routledge (2002)
[2] Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press (2014)
[3] Hutter, M.: Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer (2005)
[4] Leike, J., Hutter, M.: On the computability of AIXI. In: UAI-15. pp. 464–473. AUAI Press (2015), `http://arxiv.org/abs/1510.05572`
[5] Li, M., Vitányi, P.M.B.: An Introduction to Kolmogorov Complexity and its Applications. Springer, 3rd edn. (2008)
[6] Martin, J., Everitt, T., Hutter, M.: Death and suicide in universal artificial intelligence. Tech. rep., ANU (2016), `http://www.hutter1.net/publ/aixideath.pdf`
[7] Omohundro, S.M.: The basic AI drives. In: AGI-08. pp. 483–493. IOS Press (2008)
[8] Soares, N., Fallenstein, B., Yudkowsky, E., Armstrong, S.: Corrigibility. In: AAAI Workshop on AI and Ethics. pp. 74–82 (2015)
[9] Solomonoff, R.J.: Complexity-based induction systems: Comparisons and convergence theorems. IEEE Transactions on Information Theory IT-24, 422–432 (1978)
[10] Veness, J., Ng, K.S., Hutter, M., Uther, W., Silver, D.: A monte carlo AIXI approximation. Journal of Artificial Intelligence Research 40(1), 95–142 (2011)