Death & Suicide in Universal Artificial Intelligence

J.Martin T.Everitt M.Hutter

Artificial General Intelligence, 2016

Outline

1. Defining Death for Agents
   - Motivations
   - Agents and Environments
   - Death as a Death-state
   - Death-probability and Semimeasure Loss
2. Results
   - Known Environments: AIμ
   - Unknown Environments: AIXI
3. Conclusion

Outline

1. Defining Death for Agents
   - Motivations
   - Agents and Environments
   - Death as a Death-state
   - Death-probability and Semimeasure Loss
2. Results
   - Known Environments: AIμ
   - Unknown Environments: AIXI
3. Conclusion

Generally Intelligent Agents and Death

Why AIXI, and why agent death?

- Why do we need theoretical models of generally intelligent agents?
  - Guiding the *construction* of agents.
  - *Understanding* agent reasoning and behaviour.
  - Developing *control* strategies.
- Why study agent death?
  - AI safety and the shutdown problem.
  - Tripwire control strategies.
- Why a subjective definition of death?
  - Objective definition difficult (even for biological organisms).
  - Want to understand how the agent itself will reason about its death.

Outline

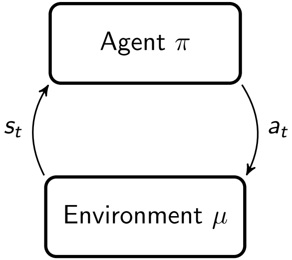
1. Defining Death for Agents
   - Motivations
   - Agents and Environments
   - Death as a Death-state
   - Death-probability and Semimeasure Loss
2. Results
   - Known Environments: AIμ
   - Unknown Environments: AIXI
3. Conclusion

The Agent-Environment Model

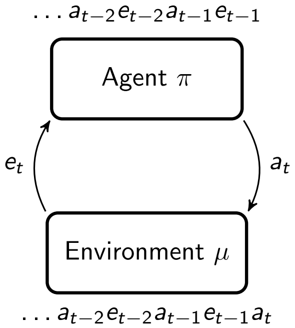
States vs. History Sequences

- Agent is a policy π: maps a history $æ_{<t}$ to an action $a_t \in \mathcal{A}$
- Environment μ: maps a history $æ_{<t}a_t$ to a percept $e_t \in \mathcal{E}$

State Model (MDP)



History Model



Australian National University

Two Generally Intelligent Agents

AIμ and AIXI

Definition (The Value Function)

The *value* (expected total future reward) of policy π in environment ν:

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

Definition (AIμ: knows the true environment)

For the true environment μ, the agent $AI\mu$ is a μ-optimal policy

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

Definition (AIXI: must learn the environment)

The agent AIXI models the environment using a mixture ξ. It is a ξ-optimal policy:

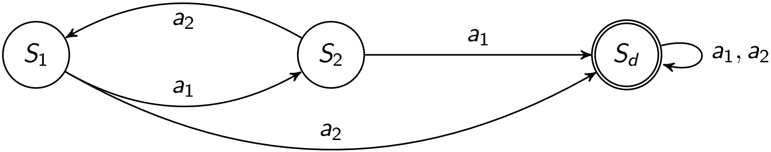
$$\pi^\xi(æ_{<t}) := \arg\max_\pi V_\xi^\pi(æ_{<t}).$$

Outline

1. Defining Death for Agents
   - Motivations
   - Agents and Environments
   - Death as a Death-state
   - Death-probability and Semimeasure Loss
2. Results
   - Known Environments: AIμ
   - Unknown Environments: AIXI
3. Conclusion

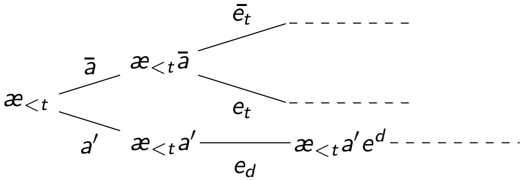
Defining a Death-State in an MDP

- In an MDP we can define a special accepting state as the death state.
- The agent remains in the death state no matter what actions it takes.



Defining a Death-State in a General Environment

- In general environments, we can't explicitly define a death *state*.
- Must instead define it via a *death-percept* $e^d \equiv (o^d, r^d)$.



Definition (Death-state in a general environment)

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

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

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.

Outline

1. Defining Death for Agents
   - Motivations
   - Agents and Environments
   - Death as a Death-state
   - Death-probability and Semimeasure Loss
2. Results
   - Known Environments: AIμ
   - Unknown Environments: AIXI
3. Conclusion

Semimeasures and Semimeasure Loss

Definition (Semimeasure)

A *semimeasure* over an alphabet $\mathcal{X}$ is a function $\nu : \mathcal{X}^* \to [0,1]$ such that

$$(1)\ \nu(\epsilon) \leq 1, \quad \text{and} \quad (2)\ \ 1 \geq \sum_{y \in \mathcal{X}} \nu(y \mid x).$$

- $\nu(x)$ is the probability that a sequence starts with the string x.
- ν may not be a proper probability measure as it need not sum to 1. There may be some probability the sequence will just terminate.

Definition (Instantaneous measure loss)

The *instantaneous measure loss* of a semimeasure ν at time t given a history $æ_{<t}a_t$ is:

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

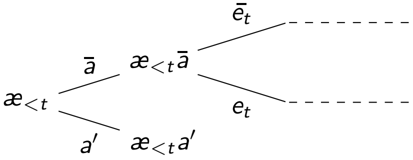
Measure Loss as Death-Probability

Definition (Semimeasure-death)

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

Advantages of this definition:

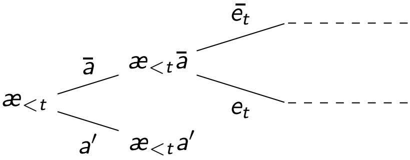
- Simple/Intuitive: No need to define a bizarre death-percept or death-state.
- General: Any sequence of death-probabilities captured by losses of some semimeasure μ.



- Equivalence of Behaviour: agents behave identically w.r.t semi-measure death and death-state.

Outline

Variance of Behaviour under Reward Range Shifts



Theorem (Self-preserving AIμ)

If rewards are bounded and non-negative, then given a history $æ_{<t}$ AIμ avoids certain immediate death:

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

Theorem (Suicidal AIμ)

If rewards are bounded and negative, then AIμ seeks certain immediate death. That is,

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

Outline

Australian National University

AIXI's Estimate of its Death-Probability

Definition (Safe and Risky Environments)

- μ is a *safe* environment if it is a proper measure with death-probability $L_\mu(æ_{<t}a_t) = 0$ for all histories $æ_{<t}a_t$. We call μ *risky* if it is not safe.
- The normalised measure μ_{norm} is thus a safe environment.

Theorem (AIXI's belief in risky environment is monotonically decreasing)

Let μ be risky s.t. $\mu \neq \mu_{\mathrm{norm}}$. Then on any history $æ_{1:t}$ the ratio of the posterior belief in μ to the posterior belief in μ_{norm} is monotonically decreasing.

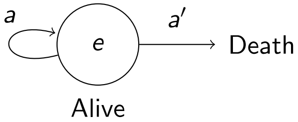
Theorem (Asymptotic ξ-probability of death in risky μ)

Let the true environment μ be computable and risky s.t. $\mu \neq \mu_{\mathrm{norm}}$. Then given any action sequence $a_{1:\infty}$, the instantaneous ξ-measure loss goes to zero w.μ.p.1 as $t \to \infty$,

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

Living Forever vs. Immortality

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



- In this 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.

Contributions

- Two definitions of Death
  - Death-State.
  - Measure Loss and Semimeasure-Death.
  - These formalisations result in identical agent behaviour.
- Known Environments: AIμ
  - Bounded Positive Rewards: AIμ avoids death.
  - Bounded Negative Rewards: AIμ seeks death.
- Unknown Environments: AIXI
  - AIXI's belief in its safety is monotonically increasing.
  - Asymptotically, AIXI's estimate of its death-probability vanishes.
  - Asypmtotically, AIXI learns it will live forever, but not that it is immortal.
- Outlook:
  - We hope this preliminary formal treatment of death will prove useful to future investigations into the shutdown problem and other problems in AI Safety related to agent termination.

Australian National University