On the Optimality of General Reinforcement Learners

Jan Leike and Marcus Hutter

http://jan.leike.name/

Australian National University

EWRL'15 — 10 July 2015

Outline

General Reinforcement Learning

How to do RL in an unknown computable (non-Markovian) environment?

General Reinforcement Learning

How to do RL in an unknown computable (non-Markovian) environment?
Idea of *AIXI* [Hut05]:

General Reinforcement Learning

How to do RL in an unknown computable (non-Markovian) environment?
Idea of *AIXI* [Hut05]:

- Disregard computation time ;)

General Reinforcement Learning

How to do RL in an unknown computable (non-Markovian) environment?
Idea of *AIXI* [Hut05]:

- Disregard computation time ;)
- Take a Bayesian mixture ξ over all computable environments

General Reinforcement Learning

How to do RL in an unknown computable (non-Markovian) environment?
Idea of *AIXI* [Hut05]:

- Disregard computation time ;)
- Take a Bayesian mixture ξ over all computable environments
- Weigh environments according to their Kolmogorov complexity

General Reinforcement Learning

How to do RL in an unknown computable (non-Markovian) environment?
Idea of *AIXI* [Hut05]:

- Disregard computation time ;)
- Take a Bayesian mixture ξ over all computable environments
- Weigh environments according to their Kolmogorov complexity
- Maximize expected rewards in the mixture

General Reinforcement Learning

How to do RL in an unknown computable (non-Markovian) environment?
Idea of *AIXI* [Hut05]:

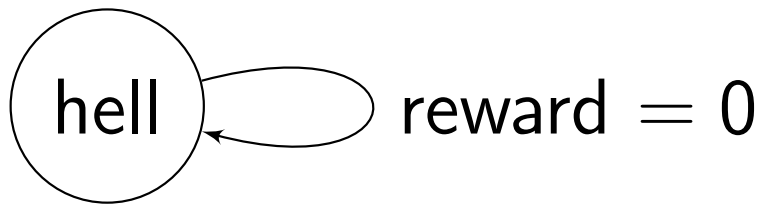
- Disregard computation time ;)
- Take a Bayesian mixture ξ over all computable environments
- Weigh environments according to their Kolmogorov complexity
- Maximize expected rewards in the mixture

Is AIXI optimal?

Outline

Hell

Hell



The Dogmatic Prior

The Dogmatic Prior

Policy π_{Lazy}:

```
while (true) { do_nothing(); }
```

The Dogmatic Prior

Policy π_{Lazy}:

```
while (true) { do_nothing(); }
```

Dogmatic prior ξ':

if not acting according to π_{Lazy},
go to hell with high probability

The Dogmatic Prior

Policy π_{Lazy}:

```
while (true) { do_nothing(); }
```

Dogmatic prior ξ':

if not acting according to π_{Lazy},
go to hell with high probability

Theorem

$AI\xi'$ acts according to π_{Lazy} as long as $V_{\xi}^{\pi_{Lazy}} > \varepsilon > 0$ (future expected reward does not get close to 0).

The Dogmatic Prior

Policy π_{Lazy}:

```
while (true) { do_nothing(); }
```

Dogmatic prior ξ':

if not acting according to π_{Lazy},
go to hell with high probability

Theorem

AIξ' acts according to π_{Lazy} as long as $V_\xi^{\pi_{Lazy}} > \varepsilon > 0$ (future expected reward does not get close to 0).

- Can be made universal ($\rightarrow$ UTM)

The Dogmatic Prior

Policy π_{Lazy}:

```
while (true) { do_nothing(); }
```

Dogmatic prior ξ':

if not acting according to π_{Lazy},
go to hell with high probability

Theorem

AIξ' acts according to π_{Lazy} as long as $V_{\xi}^{\pi_{Lazy}} > \varepsilon > 0$ (future expected reward does not get close to 0).

- Can be made universal ($\rightarrow$ UTM)
- Applies to MDPs

The Dogmatic Prior

Policy π_{Lazy}:

```
while (true) { do_nothing(); }
```

Dogmatic prior ξ':

if not acting according to π_{Lazy},
go to hell with high probability

Theorem

AIξ' acts according to π_{Lazy} as long as $V_\xi^{\pi_{Lazy}} > \varepsilon > 0$ (future expected reward does not get close to 0).

- Can be made universal ($\rightarrow$ UTM)
- Applies to MDPs
- Applies to ergodic MDPs for bounded horizon

Outline

Pareto Optimality

Pareto optimality = *no other policy Pareto-dominates my policy*

Pareto Optimality

Pareto optimality = *no other policy Pareto-dominates my policy*

Theorem ([Hut02])

AIXI is Pareto Optimal.

Pareto Optimality

Pareto optimality = *no other policy Pareto-dominates my policy*

Theorem ([Hut02])

AIXI is Pareto Optimal.

Theorem ([LH15])

Every policy is Pareto optimal.

Legg-Hutter Intelligence

Intelligence of policy π [LH07]:

$$\Upsilon_\xi(\pi) := \sum_\nu w_\nu V_\nu^\pi = V_\xi^\pi$$

Legg-Hutter Intelligence

Intelligence of policy π [LH07]:

$$\Upsilon_\xi(\pi) := \sum_\nu w_\nu V_\nu^\pi = V_\xi^\pi$$

Legg-Hutter Intelligence

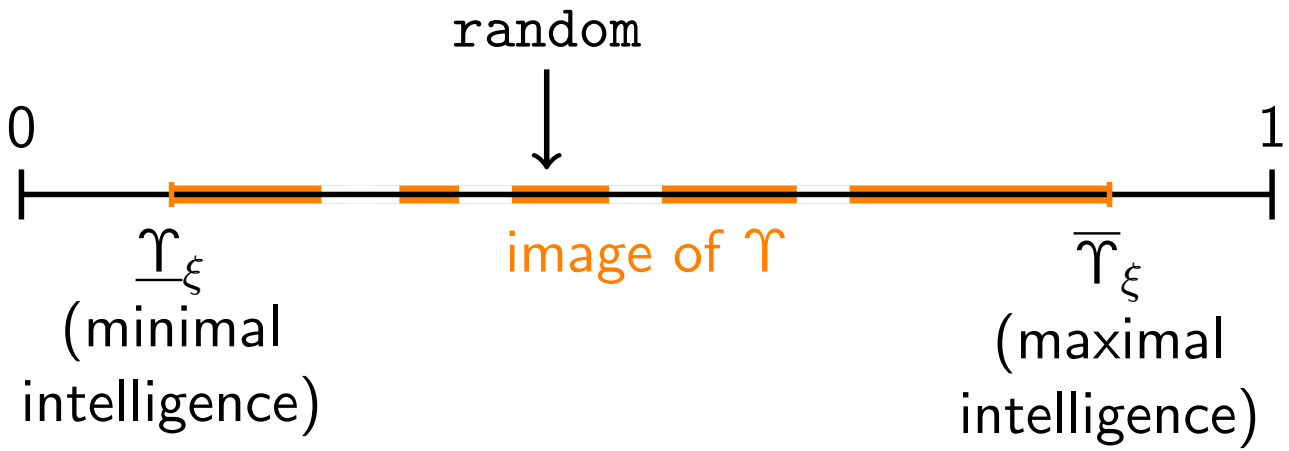
Intelligence of policy π [LH07]:

$$\Upsilon_\xi(\pi) := \sum_\nu w_\nu V_\nu^\pi = V_\xi^\pi$$

Legg-Hutter Intelligence

Intelligence of policy π [LH07]:

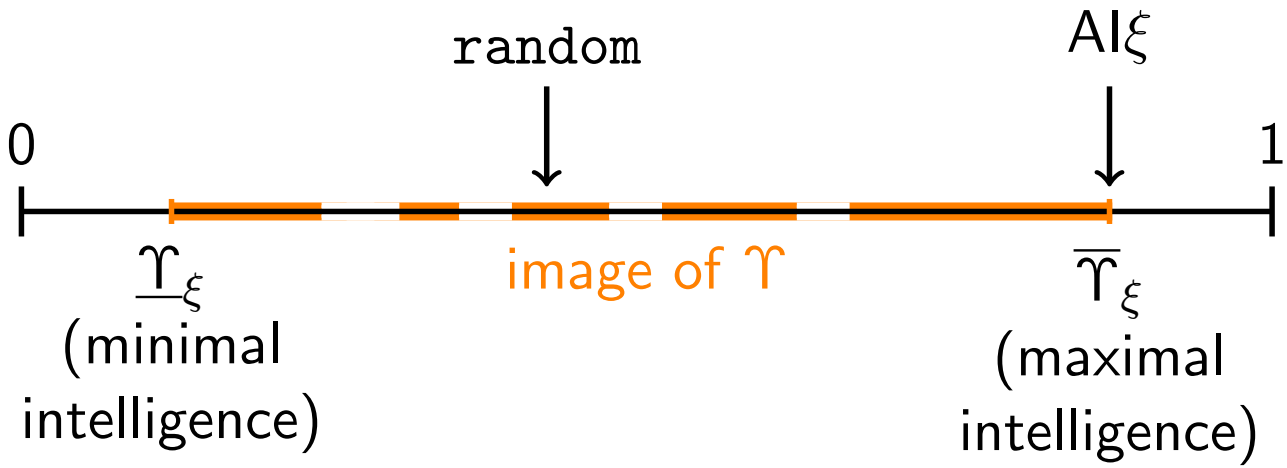
$$\Upsilon_\xi(\pi) := \sum_\nu w_\nu V_\nu^\pi = V_\xi^\pi$$



Legg-Hutter Intelligence

Intelligence of policy π [LH07]:

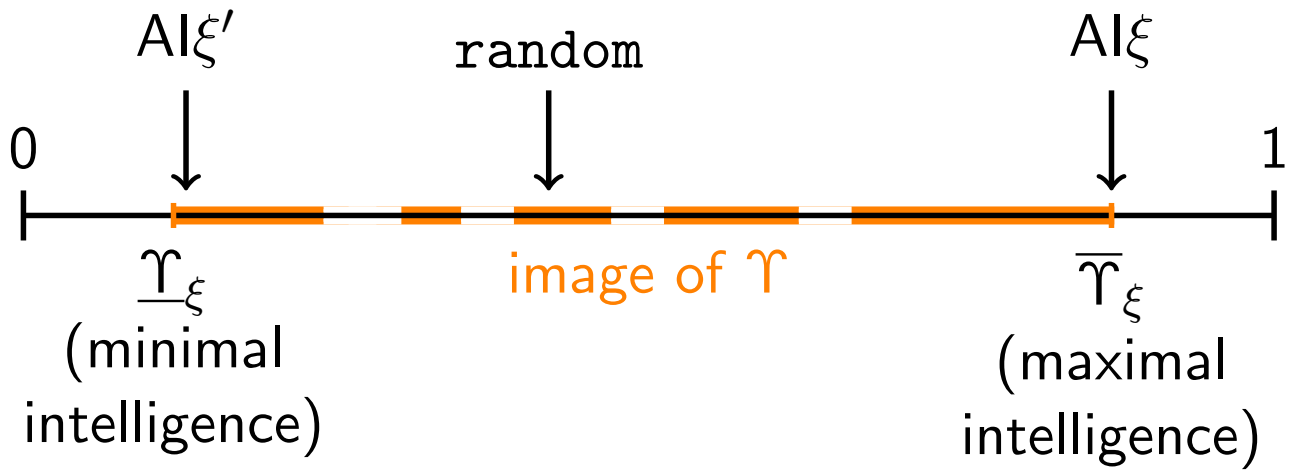
$$\Upsilon_\xi(\pi) := \sum_\nu w_\nu V_\nu^\pi = V_\xi^\pi$$



Legg-Hutter Intelligence

Intelligence of policy π [LH07]:

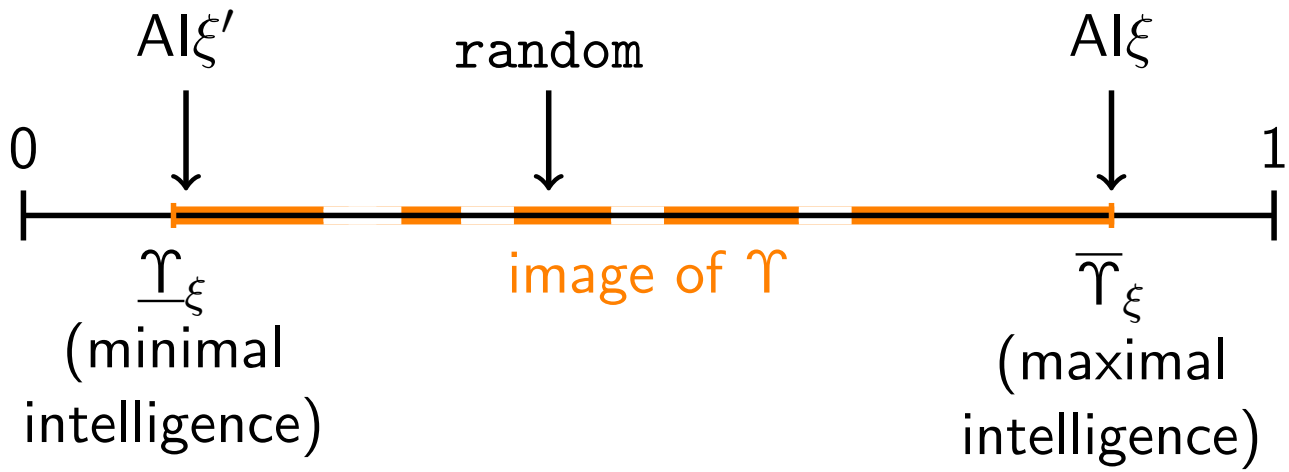
$$\Upsilon_\xi(\pi) := \sum_\nu w_\nu V_\nu^\pi = V_\xi^\pi$$



Legg-Hutter Intelligence

Intelligence of policy π [LH07]:

$$\Upsilon_\xi(\pi) := \sum_\nu w_\nu V_\nu^\pi = V_\xi^\pi$$



$\Longrightarrow$ **Legg-Hutter intelligence is highly subjective**

The Optimality of AIXI

AIXI is ...

The Optimality of AIXI

AIXI is . . .

- Pareto optimal [Hut02]

The Optimality of AIXI

AIXI is ...

- Pareto optimal [Hut02]
  Trivial: every policy is Pareto optimal

The Optimality of AIXI

AIXI is ...

- Pareto optimal [Hut02]
  Trivial: every policy is Pareto optimal
- balanced Pareto optimal [Hut02]

The Optimality of AIXI

AIXI is ...

- Pareto optimal [Hut02]
  Trivial: every policy is Pareto optimal
- balanced Pareto optimal [Hut02]
  = maximal Legg-Hutter intelligence

The Optimality of AIXI

AIXI is ...

- Pareto optimal [Hut02]
  Trivial: every policy is Pareto optimal
- balanced Pareto optimal [Hut02]
  = maximal Legg-Hutter intelligence
  highly subjective

The Optimality of AIXI

AIXI is ...

- Pareto optimal [Hut02]
  Trivial: every policy is Pareto optimal
- balanced Pareto optimal [Hut02]
  = maximal Legg-Hutter intelligence
  highly subjective
- self-optimizing [Hut02]

The Optimality of AIXI

AIXI is ...

- Pareto optimal [Hut02]
  Trivial: every policy is Pareto optimal
- balanced Pareto optimal [Hut02]
  = maximal Legg-Hutter intelligence
  highly subjective
- self-optimizing [Hut02]
  not applicable to the class of all computable environments

The Optimality of AIXI

AIXI is ...

- Pareto optimal [Hut02]
  Trivial: every policy is Pareto optimal
- balanced Pareto optimal [Hut02]
  = maximal Legg-Hutter intelligence
  highly subjective
- self-optimizing [Hut02]
  not applicable to the class of all computable environments

$\Longrightarrow$ **No formal argument for AIXI's optimality**

The Optimality of AIXI

AIXI is ...

- Pareto optimal [Hut02]
  Trivial: every policy is Pareto optimal
- balanced Pareto optimal [Hut02]
  = maximal Legg-Hutter intelligence
  highly subjective
- self-optimizing [Hut02]
  not applicable to the class of all computable environments

$$\Longrightarrow \textbf{No formal argument for AIXI's optimality}$$

Problem: Bayesian RL agents do not explore enough to lose the prior's bias

The Optimality of AIXI

AIXI is ...

- Pareto optimal [Hut02]
  Trivial: every policy is Pareto optimal
- balanced Pareto optimal [Hut02]
  = maximal Legg-Hutter intelligence
  highly subjective
- self-optimizing [Hut02]
  not applicable to the class of all computable environments

$$\Longrightarrow \textbf{No formal argument for AIXI's optimality}$$

Problem: Bayesian RL agents do not explore enough to lose the prior's bias
Solution: Add exploration through knowledge-seeking [OLH13, Lat13] or optimism [SH12]

$$\Longrightarrow \textbf{weak asymptotic optimality}$$

Does AIXI "work"?

Answer: probably, but biased by the prior

Does AIXI "work"?

Answer: probably, but biased by the prior

But: AIXI will wirehead [RO11] and then kill everyone

Outline

References I

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.

Tor Lattimore.

Theory of General Reinforcement Learning.

PhD thesis, Australian National University, 2013.

Shane Legg and Marcus Hutter.

Universal intelligence: A definition of machine intelligence.

Minds & Machines, 17(4):391–444, 2007.

References II

Jan Leike and Marcus Hutter.

Bad universal priors and notions of optimality.

In *Conference on Learning Theory*, pages 1244–1259, 2015.

Laurent Orseau, Tor Lattimore, and Marcus Hutter.

Universal knowledge-seeking agents for stochastic environments.

In *Algorithmic Learning Theory*, pages 158–172. Springer, 2013.

Mark Ring and Laurent Orseau.

Delusion, survival, and intelligent agents.

In *Artificial General Intelligence*, pages 11–20. Springer, 2011.

Peter Sunehag and Marcus Hutter.

Optimistic agents are asymptotically optimal.

In *Australasian Joint Conference on Artificial Intelligence*, pages 15–26. Springer, 2012.