General Reinforcement Learning

Jan Leike

Future of Humanity Institute
University of Oxford

9 June 2016

Reinforcement Learning Today[1]

[1] Volodymyr Mnih et al. "Human-Level Control through Deep Reinforcement Learning". In: *Nature* 518.7540 (2015), pp. 529–533.

If we upscale DQN, do we get strong AI?

Narrow Reinforcement Learning

Atari 2600	
fully observable	
ergodic	
very large state space	
ε-exploration works	

Narrow Reinforcement Learning

Atari 2600	The Real World$^{\text{TM}}$
fully observable	partially observable
ergodic	not ergodic
very large state space	infinite state space
ε-exploration works	ε-exploration fails

Narrow Reinforcement Learning

Atari 2600	The Real World^{TM}
fully observable	partially observable
ergodic	not ergodic
very large state space	infinite state space
ε-exploration works	ε-exploration fails
ergodic MDPs	general environments

Outline

AIXI

Optimality

Game Theory

AI Safety

Outline

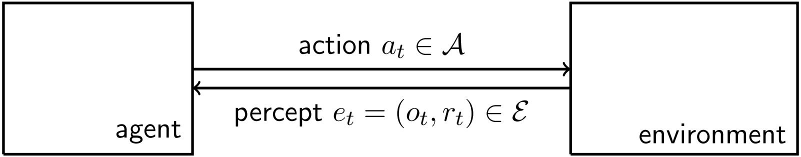
AIXI

Optimality

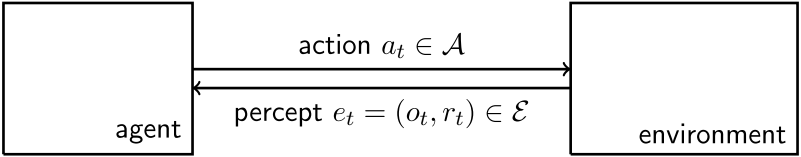
Game Theory

AI Safety

General Reinforcement Learning

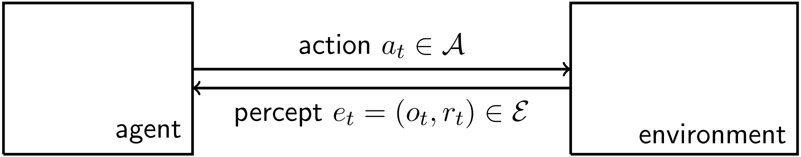


General Reinforcement Learning



policy	$\pi : (\mathcal{A} \times \mathcal{E})^* \to \Delta\mathcal{A}$
environment	$\nu : (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A} \to \Delta\mathcal{E}$
history	$æ_{<t} := a_1 e_1 a_2 e_2 \dots a_{t-1} e_{t-1}$

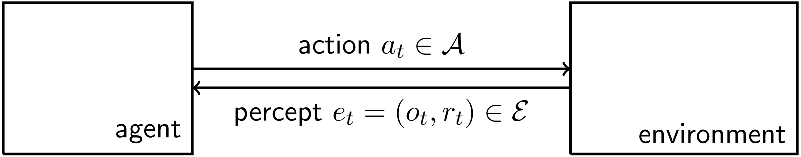
General Reinforcement Learning



policy	$\pi : (\mathcal{A} \times \mathcal{E})^* \to \Delta\mathcal{A}$
environment	$\nu : (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A} \to \Delta\mathcal{E}$
history	$æ_{<t} := a_1 e_1 a_2 e_2 \ldots a_{t-1} e_{t-1}$

Goal: maximize $\sum_{t=1}^{\infty} \gamma_t r_t$
where $\gamma : \mathbb{N} \to [0, 1]$ is a discount function with $\sum_{t=1}^{\infty} \gamma_t < \infty$

General Reinforcement Learning



policy	$\pi : (\mathcal{A} \times \mathcal{E})^* \to \Delta\mathcal{A}$
environment	$\nu : (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A} \to \Delta\mathcal{E}$
history	$æ_{<t} := a_1 e_1 a_2 e_2 \dots a_{t-1} e_{t-1}$

Goal: maximize $\sum_{t=1}^{\infty} \gamma_t r_t$
where $\gamma : \mathbb{N} \to [0, 1]$ is a discount function with $\sum_{t=1}^{\infty} \gamma_t < \infty$

Assumptions

- $0 \leq r_t \leq 1$
- $\mathcal{A}$ and $\mathcal{E}$ are finite

Value Functions

Value Functions

Value of policy π in environment ν:

$$V_\nu^\pi(æ_{<t}) := \frac{1}{\sum_{k=t}^\infty \gamma_k} \mathbb{E}_\nu^\pi \left[\sum_{k=t}^\infty \gamma_k r_k \,\middle|\, æ_{<t} \right]$$

Value Functions

Value of policy π in environment ν:

$$V_\nu^\pi(æ_{<t}) := \frac{1}{\sum_{k=t}^\infty \gamma_k} \mathbb{E}_\nu^\pi \left[\sum_{k=t}^\infty \gamma_k r_k \,\middle|\, æ_{<t} \right]$$

- optimal value: $V_\nu^* := \sup_\pi V_\nu^\pi$
- ν-optimal policy: $\pi_\nu^* := \arg\max_\pi V_\nu^\pi$

Value Functions

Value of policy π in environment ν:

$$V_\nu^\pi(æ_{<t}) := \frac{1}{\sum_{k=t}^\infty \gamma_k} \mathbb{E}_\nu^\pi \left[\sum_{k=t}^\infty \gamma_k r_k \,\middle|\, æ_{<t} \right]$$

- optimal value: $V_\nu^* := \sup_\pi V_\nu^\pi$
- ν-optimal policy: $\pi_\nu^* := \arg\max_\pi V_\nu^\pi$
- Effective horizon:

$$H_t(\varepsilon) := \min \left\{ k \,\middle|\, \frac{\sum_{i=t+k}^\infty \gamma_i}{\sum_{i=t}^\infty \gamma_i} \leq \varepsilon \right\}$$

AIXI[3]

[2] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

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

AIXI[3]

- countable set of environments $\mathcal{M} = \{\nu_1, \nu_2, \ldots\}$

[2] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

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

AIXI[3]

- countable set of environments $\mathcal{M} = \{\nu_1, \nu_2, \ldots\}$
- prior $w : \mathcal{M} \to [0, 1]$

[2] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

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

AIXI[3]

- countable set of environments $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$
- prior $w : \mathcal{M} \to [0, 1]$
  Solomonoff prior[2] $w(\nu) \propto 2^{-K(\nu)}$

[2] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

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

AIXI[3]

- countable set of environments $\mathcal{M} = \{\nu_1, \nu_2, \ldots\}$
- prior $w : \mathcal{M} \to [0, 1]$
  Solomonoff prior[2] $w(\nu) \propto 2^{-K(\nu)}$
- Bayesian mixture

$$\xi := \sum_{\nu \in \mathcal{M}} w(\nu)\nu$$

[2] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

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

AIXI[3]

- countable set of environments $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$
- prior $w : \mathcal{M} \to [0, 1]$
  Solomonoff prior[2] $w(\nu) \propto 2^{-K(\nu)}$
- Bayesian mixture

$$\xi := \sum_{\nu \in \mathcal{M}} w(\nu)\nu$$

AIXI is the Bayes-optimal agent with a Solomonoff prior

$$\pi^*_\xi := \arg\max_\pi V^\pi_\xi$$

[2] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

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

On-Policy Value Convergence

$V^{\pi}_{\xi}(æ_{<t}) - V^{\pi}_{\mu}(æ_{<t}) \to 0$ as $t \to \infty$ almost surely

Outline

AIXI

Optimality

Game Theory

AI Safety

Notions of Optimality in Reinforcement Learning

- Bayes optimality
- Asymptotic optimality
- Sample complexity bounds
- Regret bounds
- ...

Asymptotic Optimality

π is asymptotically optimal iff

$$V^*_\mu(æ_{<t}) - V^\pi_\mu(æ_{<t}) \to 0 \text{ as } t \to \infty$$

[4] Laurent Orseau. "Asymptotic Non-Learnability of Universal Agents with Computable Horizon Functions". In: *Theoretical Computer Science* 473 (2013), pp. 149–156.

Asymptotic Optimality

π is asymptotically optimal iff

$$V^*_\mu(æ_{<t}) - V^\pi_\mu(æ_{<t}) \to 0 \text{ as } t \to \infty$$

For asymptotic optimality the agent needs to explore infinitely often for an entire effective horizon.

[4] Laurent Orseau. "Asymptotic Non-Learnability of Universal Agents with Computable Horizon Functions". In: *Theoretical Computer Science* 473 (2013), pp. 149–156.

Asymptotic Optimality

π is asymptotically optimal iff

$$V^*_\mu(æ_{<t}) - V^\pi_\mu(æ_{<t}) \to 0 \text{ as } t \to \infty$$

For asymptotic optimality the agent needs to explore infinitely often for an entire effective horizon.

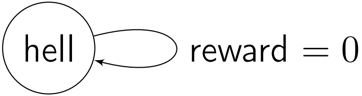
Theorem

AIXI is not asymptotically optimal.[4]

[4] Laurent Orseau. "Asymptotic Non-Learnability of Universal Agents with Computable Horizon Functions". In: *Theoretical Computer Science* 473 (2013), pp. 149–156.

Hell

Hell



The Dogmatic Prior[5]

[5] Jan Leike and Marcus Hutter. "Bad Universal Priors and Notions of Optimality". In: *Conference on Learning Theory*. 2015, pp. 1244–1259.

The Dogmatic Prior[5]

Policy π_{Lazy}:

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

[5] Jan Leike and Marcus Hutter. "Bad Universal Priors and Notions of Optimality". In: *Conference on Learning Theory*. 2015, pp. 1244–1259.

The Dogmatic Prior[5]

Policy π_{Lazy}:

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

Dogmatic prior ξ':

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

[5] Jan Leike and Marcus Hutter. "Bad Universal Priors and Notions of Optimality". In: *Conference on Learning Theory*. 2015, pp. 1244–1259.

The Dogmatic Prior[5]

Policy π_{Lazy}:

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

Dogmatic prior ξ':

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

Theorem

$\forall \varepsilon > 0 \, \exists \xi'$ s.t. AIξ' acts according to π_{Lazy} as long as $V_{\xi}^{\pi_{Lazy}}(æ_{<t}) > \varepsilon > 0$.

[5] Jan Leike and Marcus Hutter. "Bad Universal Priors and Notions of Optimality". In: *Conference on Learning Theory*. 2015, pp. 1244–1259.

Thompson Sampling

Thompson sampling policy π_T:

Sample $\rho \sim w(\cdot \mid æ_{<t})$.
Follow π^*_ρ *for* $H_t(\varepsilon_t)$ *steps.*
Repeat.

with $\varepsilon_t \to 0$.

[6] Jan Leike et al. "Thompson Sampling is Asymptotically Optimal in General Environments". In: *Uncertainty in Artificial Intelligence*. 2016.

Thompson Sampling

Thompson sampling policy π_T:

Sample $\rho \sim w(\cdot \mid æ_{<t})$*.*
Follow π^*_ρ *for* $H_t(\varepsilon_t)$ *steps.*
Repeat.

with $\varepsilon_t \to 0$.

Theorem

Thompson sampling is asymptotically optimal in mean:[6]

$$\mathbb{E}^{\pi_T}_\mu \left[V^*_\mu(æ_{<t}) - V^{\pi_T}_\mu(æ_{<t}) \right] \to 0 \text{ as } t \to \infty.$$

[6] Jan Leike et al. "Thompson Sampling is Asymptotically Optimal in General Environments". In: *Uncertainty in Artificial Intelligence*. 2016.

Recoverable Environments

An environment ν is recoverable iff

$$\sup_\pi \left| \mathbb{E}_\nu^{\pi_\nu^*}[V_\nu^*(æ_{<t})] - \mathbb{E}_\nu^\pi[V_\nu^*(æ_{<t})] \right| \to 0 \text{ as } t \to \infty.$$

Recoverable Environments

An environment ν is recoverable iff

$$\sup_{\pi} \left| \mathbb{E}_{\nu}^{\pi_{\nu}^{*}} [V_{\nu}^{*}(æ_{<t})] - \mathbb{E}_{\nu}^{\pi} [V_{\nu}^{*}(æ_{<t})] \right| \to 0 \text{ as } t \to \infty.$$

For non-recoverable environments:

Either the agent gets caught in a trap
or it is not asymptotically optimal.

Regret

$$R_m(\pi, \mu) := \max_{\pi'} \mathbb{E}_\mu^{\pi'} \left[\sum_{t=1}^{m} r_t \right] - \mathbb{E}_\mu^{\pi} \left[\sum_{t=1}^{m} r_t \right]$$

Regret

$$R_m(\pi, \mu) := \max_{\pi'} \mathbb{E}_\mu^{\pi'}\left[\sum_{t=1}^m r_t\right] - \mathbb{E}_\mu^{\pi}\left[\sum_{t=1}^m r_t\right]$$

A problem class is *learnable* iff $\exists \pi \; \forall \mu \; R_m(\pi, \mu) \in o(m)$.

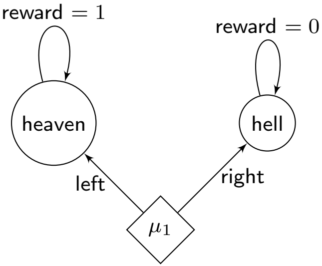
Regret

$$R_m(\pi, \mu) := \max_{\pi'} \mathbb{E}_\mu^{\pi'} \left[\sum_{t=1}^{m} r_t \right] - \mathbb{E}_\mu^{\pi} \left[\sum_{t=1}^{m} r_t \right]$$

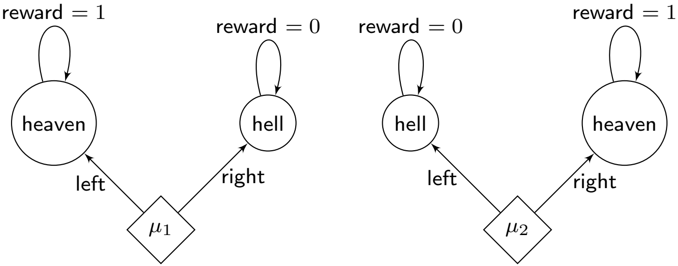
A problem class is *learnable* iff $\exists \pi \; \forall \mu \; R_m(\pi, \mu) \in o(m)$.

Fact: The general RL problem is *not* learnable.

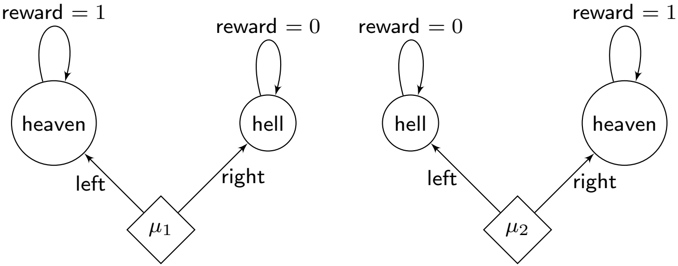
Regret in Non-Recoverable Environments



Regret in Non-Recoverable Environments



Regret in Non-Recoverable Environments



$$R_m(\text{left}, \mu_1) = 0 \qquad R_m(\text{left}, \mu_2) = m$$
$$R_m(\text{right}, \mu_1) = m \qquad R_m(\text{right}, \mu_2) = 0$$

Sublinear Regret

Theorem

If

- $\mu \in \mathcal{M}$ *is recoverable,*
- π *is asymptotically optimal in mean, and*
- γ *satisfies some weak assumptions,*

then regret is sublinear.[7]

[7] Jan Leike et al. "Thompson Sampling is Asymptotically Optimal in General Environments". In: *Uncertainty in Artificial Intelligence*. 2016.

Optimality Summary

	AIXI	TS	All policies
Sublinear regret	×	recoverable	×
Sample complexity	×	?	
Pareto optimality	✓	✓	✓
Bayes optimality	✓	×	
Asymptotic optimality	×	✓	

Outline

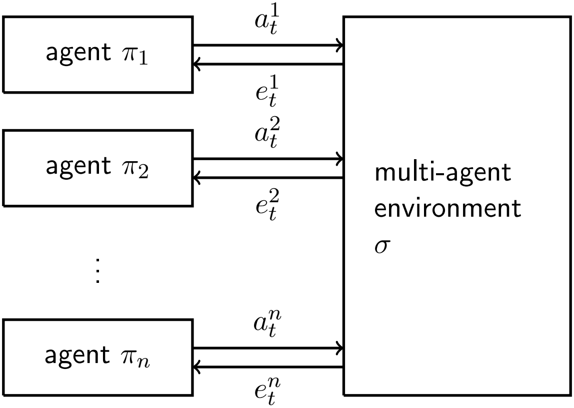
AIXI

Optimality

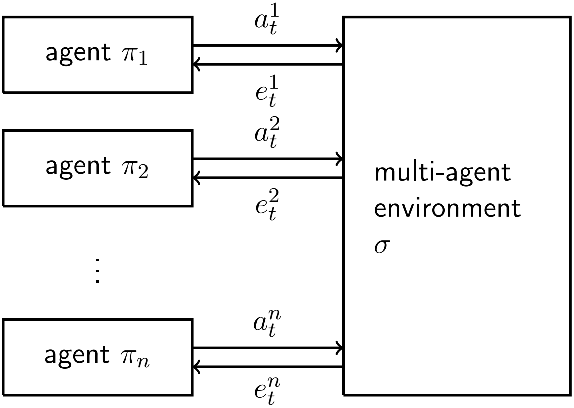
Game Theory

AI Safety

Multi-Agent Environments

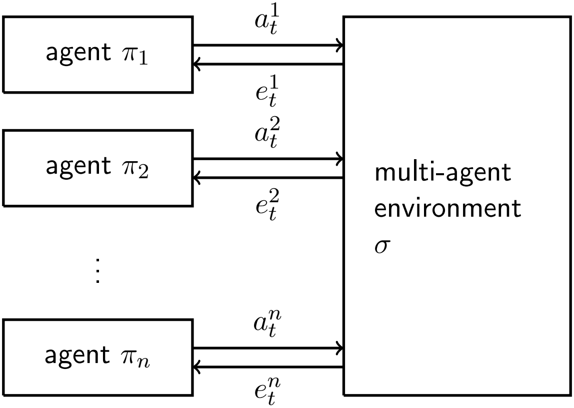


Multi-Agent Environments



- π_i is an ε-best response iff $V^*_{\sigma_i} - V^{\pi_i}_{\sigma_i} < \varepsilon$

Multi-Agent Environments



- π_i is an ε-best response iff $V^*_{\sigma_i} - V^{\pi_i}_{\sigma_i} < \varepsilon$
- $\pi_1, \ldots, \pi_n$ play an ε-Nash equilibrium iff each π_i is an ε-best response

The Bayesian Approach

- countable set of policies Π

The Bayesian Approach

- countable set of policies Π
- prior $w \in \Delta\Pi$

The Bayesian Approach

- countable set of policies Π
- prior $w \in \Delta\Pi$
- act Bayes-optimal with respect to w

The Bayesian Approach

- countable set of policies Π
- prior $w \in \Delta\Pi$
- act Bayes-optimal with respect to w

Grain of Truth: the Bayes-optimal policy needs to be in Π

Results for Bayesian Agents

Theorem

If each player is Bayesian, knows the infinite repeated game and has a grain of truth, then the players converge to an ε-Nash equilibrium.[8]

[8] Ehud Kalai and Ehud Lehrer. "Rational Learning Leads to Nash Equilibrium". In: *Econometrica* (1993), pp. 1019–1045.

[9] Jan Leike, Jessica Taylor, and Benya Fallenstein. "A Formal Solution to the Grain of Truth Problem". In: *Uncertainty in Artificial Intelligence*. 2016.

Results for Bayesian Agents

Theorem

If each player is Bayesian, knows the infinite repeated game and has a grain of truth, then the players converge to an ε-Nash equilibrium.[8]

Theorem

Two Bayesian players playing infinite repeated matching pennies may fail to converge to an ε-Nash equilibrium, even if they have a grain of truth.[9]

---

[8] Ehud Kalai and Ehud Lehrer. "Rational Learning Leads to Nash Equilibrium". In: *Econometrica* (1993), pp. 1019–1045.

[9] Jan Leike, Jessica Taylor, and Benya Fallenstein. "A Formal Solution to the Grain of Truth Problem". In: *Uncertainty in Artificial Intelligence*. 2016.

Solving the Grain of Truth Problem[10]

Theorem

There is a class of environments $\mathcal{M}_{refl}$ that contains a grain of truth with respect to any computable priors' Bayes-optimal policies in any computable multi-agent environment.

[10] Jan Leike, Jessica Taylor, and Benya Fallenstein. "A Formal Solution to the Grain of Truth Problem". In: *Uncertainty in Artificial Intelligence*. 2016.

Solving the Grain of Truth Problem[10]

Theorem

There is a class of environments $\mathcal{M}_{refl}$ that contains a grain of truth with respect to any computable priors' Bayes-optimal policies in any computable multi-agent environment.

Theorem

Each $\nu \in \mathcal{M}_{refl}$ is limit computable.

[10] Jan Leike, Jessica Taylor, and Benya Fallenstein. "A Formal Solution to the Grain of Truth Problem". In: *Uncertainty in Artificial Intelligence*. 2016.

Solving the Grain of Truth Problem[10]

Theorem
There is a class of environments $\mathcal{M}_{refl}$ that contains a grain of truth with respect to any computable priors' Bayes-optimal policies in any computable multi-agent environment.

Theorem
Each $\nu \in \mathcal{M}_{refl}$ is limit computable.

Theorem
There are limit computable policies $\pi_1, \ldots, \pi_n$ such that for any computable multi-agent environment σ and for all $\varepsilon > 0$ and all $i \in \{1, \ldots, n\}$ the probability that the policy π_i is an ε-best response converges to 1 as $t \to \infty$.

[10] Jan Leike, Jessica Taylor, and Benya Fallenstein. "A Formal Solution to the Grain of Truth Problem". In: *Uncertainty in Artificial Intelligence*. 2016.

Outline

AIXI

Optimality

Game Theory

AI Safety

AI Safety Approaches

bottom-up	top-down
practical algorithms	
toy models	
demos	

AI Safety Approaches

bottom-up	top-down
practical algorithms	theoretical models
toy models	abstract problems
demos	theorems

"Applications" of GRL to AI Safety

- self-modification: Orseau and Ring (2011), Orseau and Ring (2012), Everitt et al. (2016)
- self-reflection: Fallenstein, Soares, and Taylor (2015), Leike, Taylor, and Fallenstein (2016)
- memory manipulation: Orseau and Ring (2012)
- interruptibility: Orseau and Armstrong (2016)
- decision theory: Everitt, Leike, and Hutter (2015)
- wireheading: Ring and Orseau (2011), Everitt and Hutter (2016)
- value learning: Dewey (2011)
- questions of identity: Orseau (2014)

Limits of the Current Model

- model-based
- dualistic
- not self-improving
- assumes infinite computation

Conclusion

Mathematical and mental tools to think about strong AI

- exploration vs. exploitation
- effective horizon
- on-policy vs. off-policy
- model-based vs. model-free
- recoverability
- asymptotic optimality
- reflective oracles

Jan Leike. "Nonparametric General Reinforcement Learning". PhD thesis. Australian National University, 2016

http://jan.leike.name/