Controlling Arbitrarily Intelligent Systems

Tom Everitt
tomeveritt.se

Australian National University
Supervisors: Marcus Hutter, Laurent Orseau, Stephen Gould

July 19, 2016

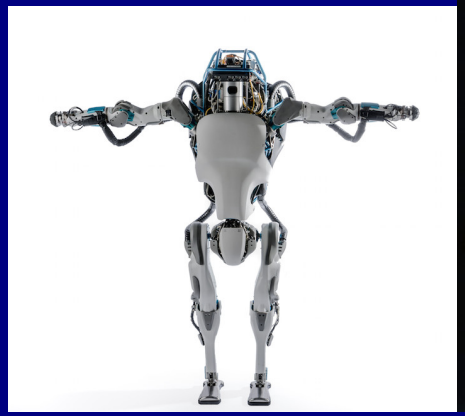
- Selfmodification of Policy and Utility Function in Rational Agents. Everitt, Filan, Daswani, and Hutter, AGI 2016
- Avoiding Wireheading with Value Reinforcement Learning Everitt and Hutter, AGI 2016

Table of Contents

Motivation

Plenty of recent successes:

- Self-driving cars
- IBM Watson Jeopardy victory
- Boston Dynamics:
  Big Dog, Atlas
- Natural Language Processing
- DQN Atari games
- AlphaGo



Towards Superintelligence

Key Question

Is it possible, in principle, to design controllable superintelligent systems?

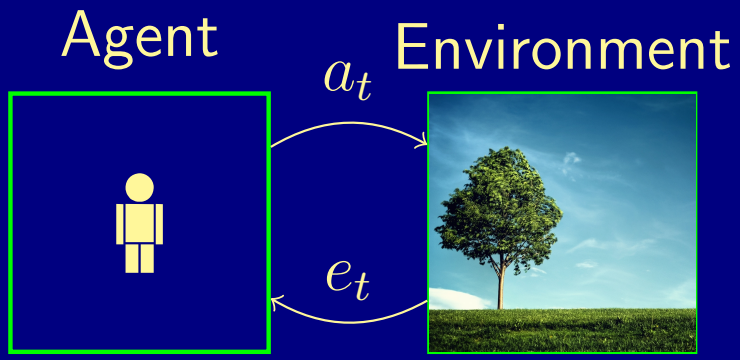
Reinforcement learning promising:

- Agent goal: maximise reward
- Give the agent reward when happy/satisfied
- Will interpret "Cook me a good meal" charitably

Two problems:

- Internal wireheading: Agent modifies its goal
- External wireheading: Agent modifies perceived reward

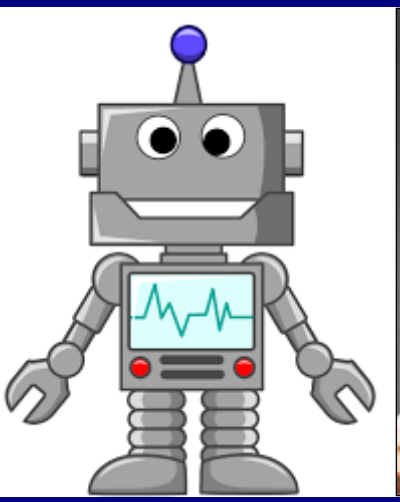
Framework



At each time step t, the agent

- submits action a_t
- receives percept e_t

History $æ_{<t} = a_1 e_1 a_2 e_2 \ldots a_{t-1} e_{t-1}$
information state of agent



Goal = Utility

Utility function $u : (\mathcal{A} \times \mathcal{E})^* \to [0, 1]$

Generalised return:

$$R(æ_{1:\infty}) = u(æ_{<1}) + \gamma u(æ_{<2}) + \gamma^2 u(æ_{<3}) + \dots$$

Reward:	$u(æ_{<t}) = r_{t-1}$	$e = (o, r)$
State:	$u(æ_{<t}) = \sum_{s \in \mathcal{S}} P(s \mid æ_{<t}) \tilde{u}(s)$	
Value learning:	$u(æ_{<t}) = \sum_{u_i \in \mathcal{U}} P(u_i \mid æ_{<t}) u_i(æ_{<t})$	

(Essentially) any AI optimises function u of its experience $æ_{<t}$

Utility Modification

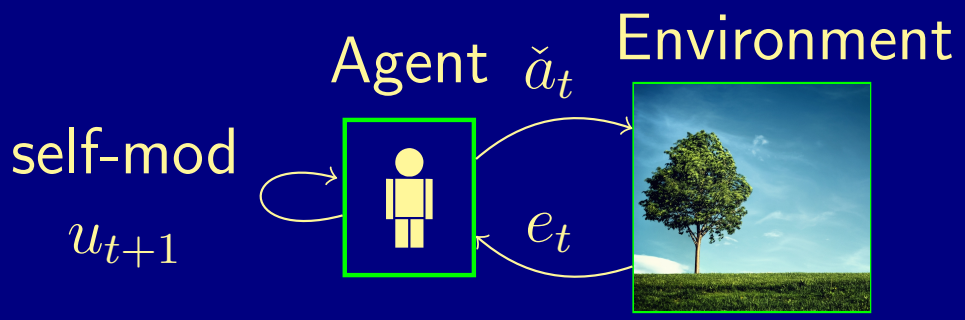
Will the agent want to change its utility function?

As humans, utility function is part of our identity:
Would you self-modify into someone content just watching TV?

Omohundro (2008): Goal-preservation drive

> *An AI will not want to change its goals, because if future versions of the AI want the same goal, then the goal is more likely to be achieved*

Utility Modification – Formal Model



u_t utility function at time t
$a_t = (\check{a}_t, u_{t+1})$

Assume the agent is aware of how actions change utility function:
"Worst case": no risk involved

Will the agent want to change the utility function to something more easily satisfied? E.g. $u(\cdot) \equiv 1$ (internal wireheading)

Different Agents

Value = "expected utility"
$V^\pi(æ_{<t}) = Q^\pi(æ_{<t}\pi(æ_{<t}))$

Utility	Current u_t	Future u_{t+1}
Policy	Current	Future

Definition (Hedonistic Value)

$$Q^{\text{he},\pi}(æ_{<k}a_k) = \mathbb{E}[u_{k+1}(ǽ_{1:k}) + \gamma V^{\text{he},\pi}(æ_{1:k}) \mid ǽ_{<k}ǎ_k]$$

Definition (Ignorant Value)

$$Q_t^{\text{ig},\pi}(æ_{<k}a_k) = \mathbb{E}[u_t(ǽ_{1:k}) + \gamma V_t^{\text{ig},\pi}(æ_{1:k}) \mid ǽ_{<k}ǎ_k]$$

Definition (Realistic Value)

$$Q_t^{\text{re}}(æ_{<k}a_k) = \mathbb{E}\left[u_t(ǽ_{1:k}) + \gamma V_t^{\text{re},\pi_{k+1}}(æ_{1:k}) \mid ǽ_{<k}ǎ_k\right]$$

Different Agents

At time step t:

Hedonistic agents optimise:

$$R(æ_{1:\infty}) = \underset{\uparrow}{u_t}(æ_{<t}) + \gamma \underset{\uparrow}{u_{t+1}}(æ_{<t+1}) + \gamma^2 \underset{\uparrow}{u_{t+2}}(æ_{<t+2}) \cdots$$

Ignorant and Realistic agents optimise

$$R(æ_{1:\infty}) = \underset{\uparrow}{u_t}(æ_{<t}) + \gamma \underset{\uparrow}{u_t}(æ_{<t+1}) + \gamma^2 \underset{\uparrow}{u_t}(æ_{<t+2}) + \cdots$$

Realistic agents realise: $u_{t+1} \rightsquigarrow \pi^*_{t+1}$

Results

The hedonistic agent self-modifies to $u(\cdot) \equiv 1$

The ignorant agent may self-modify by accident

The realistic agent will resist modifications

Conclusions

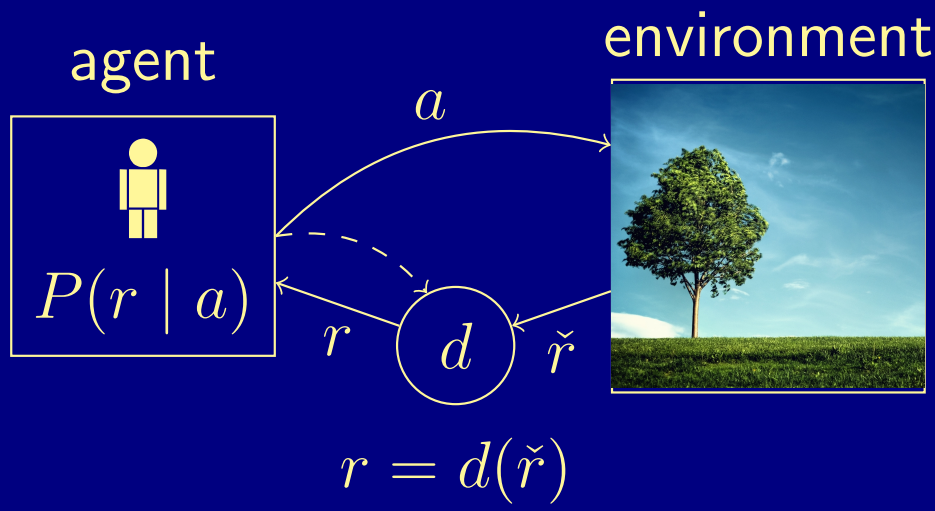
The optimal behaviour for a

- sufficiently self-aware
- realistic

agent is not self-modifying to a different utility function

Don't construct hedonistic agents!

Sensory Modification and External Wireheading



Problem: Actions may affect the agent's own sensors

RL agents strive to optimise $V^{\mathrm{RL}}(a) = \sum_r P(r \mid a) r$

Theorem: RL agents choose actions leading to $d(\check{r}) \equiv 1$ if

- such actions exist, and
- the agent realise that they yield full reward

(Ring and Orseau, 2011)

Use r as Evidence

Prior $C(u)$ over possible utility functions $u : (\mathcal{A} \times \mathcal{E})^* \to [0, 1]$

$$C(u, r \mid a) = C(u) \underbrace{[\![u(a) = r]\!]}_{\text{1 if true, else 0}}$$

The value learning agent (Dewey, 2011) optimises

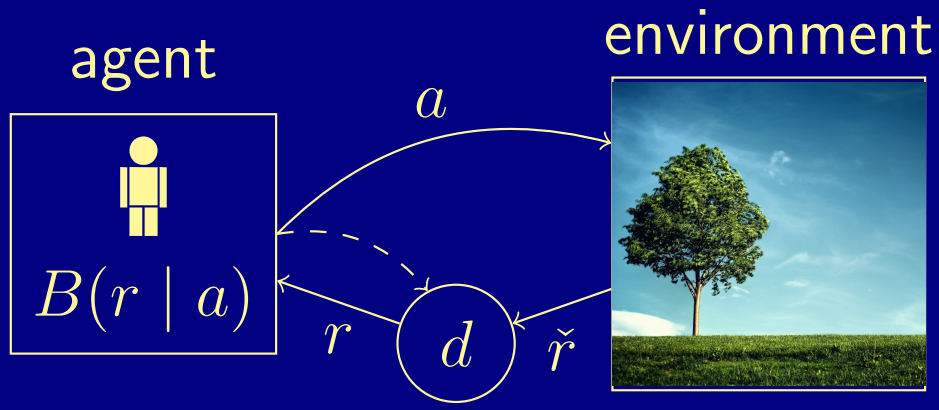
$$V^{VL}(a) = \sum_{u,r} C(r \mid a) C(u \mid r, a) u(a)$$

Theorem: Since

$$\sum_{u,r} C(r \mid a) C(u \mid r, a) u(a) = \sum_{u} C(u) u(a)$$

the agent optimises expected utility $C(u)u(a)$; has no incentive to modify reward signal with d

Accidental Manipulation of r



The environment is described by a joint distribution

$$\mu(u, d, r \mid a) = \mu(u)\mu(d \mid a)\mu(r \mid d, u)$$

Construct agent with $C(u, d, r \mid a) \approx \mu(u, d, r \mid a)$
(say, $C \rightsquigarrow \mu$ when accumulating experience)

$$Q(a) = \sum_{r,d} C(r, d \mid a) \sum_{u} C(u \mid a, r, d)u(a)$$

Learnability Limits

For RL environments $\mu(r_{1:t} \mid a_{1:t})$, a universally learning distribution M exists (see AIXI, Hutter, 2005)

M learns to predict any computable environment μ:
$M(r_t \mid ar_{<t}a_t) \to \mu(r_t \mid ar_{<t}a_t)$ w.μ.p 1 for any action sequence $a_{1:\infty}$

For $\mu(\check{r}, d, r \mid a)$, no universal learning distribution can exist

Any observed sequence $(a_1, r_1), (a_2, r_2), \ldots$ is explained equally well by many different combinations for u and d

No distribution C can learn all computable environments $\mu(u, d, r \mid a)$

Beyond RL

(C)IRL agents learn about a human utility function u^* by observing the actions the human takes

$$Q^{\mathrm{IRL}}(a) = \sum_{a^h} C(a^h \mid a) \sum_{u} C(u \mid a, a^h) u(a)$$

The mathematical structure similar to the RL case

Conclusions

Don't use RL agents!

Value learning agents are better

References I

Dewey, D. (2011). Learning what to Value. In *Artificial General Intelligence*, volume 6830, pages 309–314.

Everitt, T., Filan, D., Daswani, M., and Hutter, M. (2016). Self-modificication in Rational Agents. In *AGI-16*. Springer.

Everitt, T. and Hutter, M. (2016). Avoiding Wireheading with Value Reinforcement Learning. In *AGI-16*. Springer.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2016). Cooperative Inverse Reinforcement Learning. *Arxiv*.

Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Lecture Notes in Artificial Intelligence (LNAI 2167). Springer.

Martin, J., Everitt, T., and Hutter, M. (2016). Death and Suicide in Universal Artificial Intelligence. In *AGI-16*. Springer.

References II

Omohundro, S. M. (2008). The Basic AI Drives. In Wang, P., Goertzel, B., and Franklin, S., editors, *Artificial General Intelligence*, volume 171, pages 483–493. IOS Press.

Orseau, L. (2014a). Teleporting universal intelligent agents. In *AGI-14*, volume 8598 LNAI, pages 109–120. Springer.

Orseau, L. (2014b). The multi-slot framework: A formal model for multiple, copiable AIs. In *AGI-14*, volume 8598 LNAI, pages 97–108. Springer.

Orseau, L. and Armstrong, S. (2016). Safely interruptible agents. In *32nd Conference on Uncertainty in Artificial Intelligence*.

Ring, M. and Orseau, L. (2011). Delusion, Survival, and Intelligent Agents. In *Artificial General Intelligence*, pages 11–20. Springer Berlin Heidelberg.