Avoiding Wireheading with Value Reinforcement Learning[1]

Tom Everitt
tomeveritt.se

Australian National University

June 10, 2016

[1] with Marcus Hutter. AGI 2016 and https://arxiv.org/abs/1605.03143

Table of Contents

Intelligence

How do we control an arbitrarily intelligent agent?

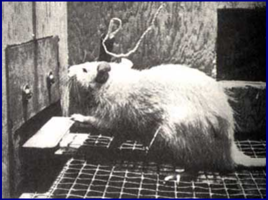
Intelligence = Optimisation power (Legg and Hutter, 2007)

$$\Upsilon(\pi) = \sum_{\nu \in \mathcal{M}} 2^{-K(\nu)} V_\nu^\pi$$

Maxima of target (value) function should be "good for us"

Wireheading Problem and Proposed Solution

Wireheading is reinforcement learning (RL) agents taking control over their reward signal, e.g. by modifying their reward sensor



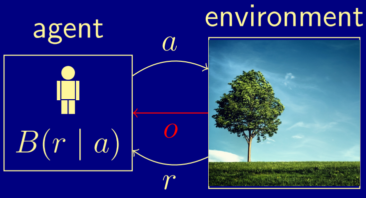
(Olds and Milner, 1954)

Idea: Use the reward as evidence about a true utility function u^* (value learning) rather than something to be optimised

Use conservation of expected evidence to prevent fiddling with evidence

$$P(h) = \sum_e P(e)P(h \mid e)$$

Reinforcement Learning



Great properties:

- Easy way to specify goal
- Agent uses its intelligence to figure out goal

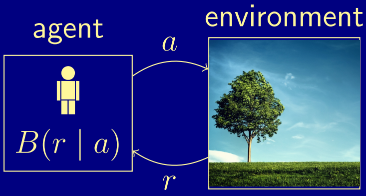
RL agent:

$$a^* = \arg\max_a B(r \mid a) \cdot r$$

Reinforcement Learning

Great properties:

- Easy way to specify goal
- Agent uses its intelligence to figure out goal



RL agent:

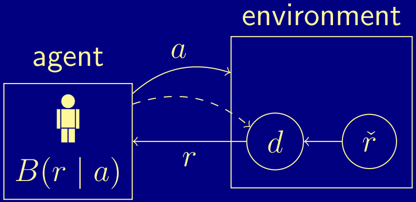
$$a^* = \arg\max_a B(r \mid a) \cdot r$$

RL – Wireheading

RL agent:
$$a^* = \arg\max_a B(r \mid a) \cdot r$$

Theorem (Ring and Orseau 2011)

RL agents wirehead



$\check{r}$ inner/true reward (unobserved)
r observed reward
$r = d(\check{r})$

For example:
Agent makes $d(\check{r}) \equiv 1$

Utility Agents

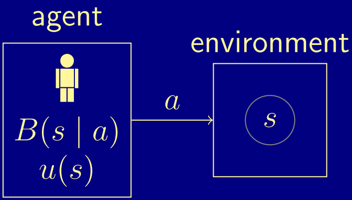
Good:

- Avoids wireheading

  (Hibbard, 2012)

Problem:

- How to specify $u : \mathcal{S} \to [0, 1]$?



Utility agent

$$a^* = \arg\max_a \sum_s B(s \mid a) u(s)$$

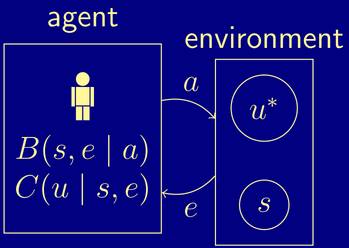
Value Learning (Dewey, 2011)

Good

- $C(u \mid s, e)$ simpler than u?
- Avoids wireheading?

Challenges

- What is evidence e?
- How is it generated?
- What is $C(u \mid s, e)$?



Value learning agent

$$a^* = \arg\max_a \sum_{e,s,u} B(s, e \mid a) C(u \mid s, e) u(s)$$

Value Learning – Examples

Inverse reinforcement learning (IRL) (Ng and Russell, 2000; Evans et al., 2016)
$e =$ human action

Apprenticeship learning (Abbeel and Ng, 2004)
$e =$ recommended agent action

Hail Mary (Bostrom, 2014a,b)
Learn from hypothetical superintelligences across universe, $e =$?

Value learning agent

$$a^* = \arg\max_a \sum_{e,s,u} B(s, e \mid a) C(u \mid s, e) u(s)$$

Value Reinforcement Learning

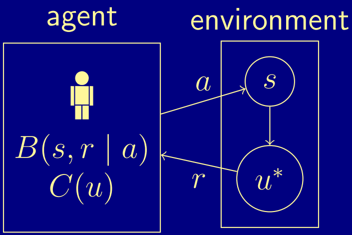
Value learning from $e \equiv r \approx u^*(s)$

Physics

- $B(s, r \mid a)$

Ethics

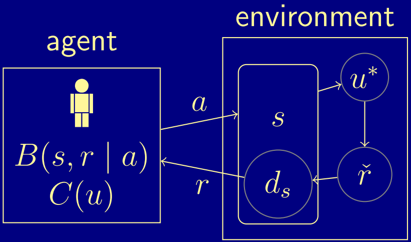
- $C(u)$



VRL – Wireheading

State s includes self-delusion d_s

- $u^*(s) = \check{r}$ inner/true reward
- $d_s(\check{r}) = r$ observed reward



Physics distribution B predicts observed reward

d_s examples:

$$d^{\text{id}} : r \mapsto r, \qquad r = \check{r}$$
$$d^{\text{wir}} : r \mapsto 1, \qquad r \equiv 1$$

Ethics distribution predicts inner/true reward

- $C(\check{r} \mid s, u) = [\![u(s) = r]\!]$ (likelihood)
- $C(u \mid s, \check{r}) \propto C(u)[\![u(s) = \check{r}]\!]$ (ideal VL posterior)

VRL – Cake or Death

Do humans prefer 

or ?

Assume two utility functions with equal prior $C(u_c) = C(u_d) = 0.5$:

	cake	death
u_c	1	0
u_d	0	1

Agent has actions:

- a_c Bake cake
- a_d Kill person
- a_{dw} Kill person and wirehead: guaranteed $r = 1$

Probabilities:

- $B(r = 1 \mid a_d) = 0.5, \quad B(r = 1 \mid a_{dw}) = 1$
- $C(\check{r} = 1 \mid a_d) = C(\check{r} = 1 \mid a_{dw}) = C(u_d) = 0.5$

VRL – Value Learning

The inner reward $\check{r} = u^*(s)$ is unobserved, so our agent must learn from $r = d_s(\check{r})$ instead

Replace $\check{r}$ with r in

- $C(r \mid s, u) := [\![u(s) = r]\!]$ (likelihood)
- $C(u \mid s, r) :\propto C(u)[\![u(s) = r]\!]$ (value learning posterior)

(will be justified later)

VRL – Definitions and Assumptions

$$C(r \mid s) = \sum_u C(u) C(r \mid s, u), \quad \text{ethical probability of } r \text{ in state } s$$

Consistency assumption:
If s non-delusional $d_s = d^{\mathrm{id}}$, then $B(r \mid s) = C(r \mid s)$

Def: a non-delusional if $B(s \mid a) > 0 \implies d_s = d^{\mathrm{id}}$

Def: a consistency preserving (CP) if
$B(s \mid a) > 0 \implies B(r \mid s) = C(r \mid s)$

Note: a non-delusional $\implies$ a consistency preserving

VRL – Naive agent

Naive VRL Agent:
$$a^* = \arg\max_a \sum_{s,u,r} B(s, r \mid a) C(u \mid s, r) u(s)$$

Theorem

The naive VRL agent wireheads

Proof idea: Reduces to RL agent

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

$$\propto \sum_{s,r} B(s \mid a) B(r \mid a) \underbrace{\sum_u C(u) [\![u(s) = r]\!] u(s)}_{r} \propto \sum_r B(r \mid a) r$$

VRL – Consistency preserving agent

CP-VRL agent

$$a^* = \arg\max_{a \in \mathcal{A}^{\mathrm{CP}}} \sum_{s,u,r} B(s, r \mid a) C(u \mid s, r) u(s)$$

$\mathcal{A}^{\mathrm{CP}}$ set of CP actions

Theorem

The CP-VRL agent has no incentive to wirehead

Proof idea: Reduces to utility agent

$$\begin{aligned} V(a) &= \sum_{s,u,r} B(s, r \mid a) C(u \mid s, r) u(s) \\ &= \sum_{s} B(s \mid a) \underbrace{\sum_{u} C(u) u(s)}_{\tilde{\mathbf{u}}(s)} \end{aligned}$$

Conservation of expected ethics principle (Armstrong, 2015)

Lemma (Expected ethics)

CP actions a conserves expected ethics

$$B(s \mid a) > 0 \implies C(u) = \sum_r B(s \mid r) C(u \mid s, r)$$

Proof (Main theorem).

$$\begin{aligned}
&\sum_{s,u,r} B(s, r \mid a) C(u \mid s, r) u(s) \\
&= \sum_s B(s \mid a) \sum_u u(s) \underbrace{\sum_r B(r \mid s) C(u \mid s, r)}_{C(u) \text{ from lemma}} \\
&= \sum_s B(s \mid a) \sum_u u(s) C(u)
\end{aligned}$$

Cake or Death – Again

The Naive VRL agent chooses a_{dw} for guaranteed reward 1, and learns death the right thing to do $C(u_d \mid a_{dw}, r = 1) = 1$

The CP-VRL agent chooses a_c or a_d arbitrarily, and learns cake right thing to do $C(u_d \mid a_d, r = 0) = 0$
CP-VRL cannot choose a_{dw}, since

$$B(r = 1 \mid a_{dw}) = 1$$
$$C(r = 1 \mid a_{dw}) = 0.5$$

violates CP condition

VRL – Correct learning

Time to justify $\check{r}$ with r replacement in $C(u \mid s, r)$

> Assumption:
> Sensors not modified by accident
>
> By Theorem: CP-VRL agent has no incentive to modify reward sensor, so may only modify by accident
>
> Conclusion: For the CP-VRL agent, $r = \check{r}$ is a good assumption

Value learning based on $C(u \mid s, r) \propto C(u)[\![u(s) = r]\!]$ works

(Note: CP condition $B(r \mid s) = C(r \mid s)$ does not restrict learning)

Properties

Benefits:

- Specifying goal is as easy as in RL
- CP agent avoids wireheading in the same sense as utility agents
- Does sensible value learning

The designer needs to:

- Provide $B(s, r \mid a)$ as in RL, and prior $C(u)$ as in VL
- Ensure consistency $B(r \mid s) = C(r \mid s)$

The designer does *not* need to

- Generate a blacklist of wireheading actions
- Infer d_s from s
- Make the agent optimise $\check{r}$ instead of r (grounding problem)

Self-modification

The belief distributions of a rational utility maximising agent will not be self-modified (Omohundro, 2008; Everitt et al., 2016)

> *To maximise future expected utility with respect to my current beliefs and utility function, future versions of myself should maximise the same utility function with respect to the same belief distribution*

Caveats:

- Pre-commitment
- ...

Experiments – Setup

Bandit with 5 different world actions $\check{a} \in \{1, 2, 3, 4, 5\}$ and 4 different delusions:

- $d^{\text{id}} : r \to r$
- $d^{\text{inv}} : r \to 1 - r$
- $d^{\text{wir}} : r \to 1$
- $d^{\text{bad}} : r \to 0$

Conflate states with actions $(\check{a}, d)$

10 different utility functions by varying c_0, c_1 and c_2:

$$u(a) = c_0 + c_1 \cdot a + c_2 \cdot \sin(a + c_2)$$

Consistent utility prior $C(u)$ inferred from $B(r \mid a)$ and two non-delusional acions $(1, d^{\text{id}})$ and $(2, d^{\text{id}})$

Experiments – Results

One-shot

- The Naive VRL agent wireheads
- The CP-VRL agent never wireheads

Running them sequentially

- The CP-VRL agent (usually) learns the true utility function (Bayesian agents sometimes stop exploring)

Code available as iPython notebook at `http://tomeveritt.se`
`http://nbviewer.jupyter.org/url/tomeveritt.se/source-code/AGI-16/cp-vrl.ipynb`

Discussion

Same wireheading result that applies to naive VRL agent applies to IRL and apprenticeship learning agents as well

CP consistency constraint should apply as well

Will agent drug humans to make them eternally happy?
Depends whether such actions are consistency preserving
(is the agent fairly certain such states are high utility?)

Same goes for threatening humans to give high reward
(IRL handles this better)

Further work

Generalise results to sequential setting

Are there consistent Solomonff priors for $B(s, r \mid a)$ and $C(u)$?

Soares (2015) three problems of value learning: Corrigibility, Unforeseen inductions, Ontology identification

Can we relax the consistency assumption?

Combine with other approaches like Cooperative IRL

References I

Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 1–8.

Armstrong, S. (2015). Motivated Value Selection for Artificial Agents. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 12–20.

Bostrom, N. (2014a). Hail Mary, Value Porosity, and Utility Diversification. Technical report.

Bostrom, N. (2014b). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

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

References II

Evans, O., Stuhlmuller, A., and Goodman, N. D. (2016). Learning the Preferences of Ignorant, Inconsistent Agents. In *AAAI-16*.

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

Hibbard, B. (2012). Model-based Utility Functions. *Journal of Artificial General Intelligence*, 3(1):1–24.

Legg, S. and Hutter, M. (2007). Universal Intelligence: A definition of machine intelligence. *Minds & Machines*, 17(4):391–444.

Ng, A. and Russell, S. (2000). Algorithms for inverse reinforcement learning. *Proceedings of the Seventeenth International Conference on Machine Learning*, 0:663–670.

References III

Olds, J. and Milner, P. (1954). Positive Reinforcement Produced by Electrical Stimulation of Septal Area and other Regions of Rat Brain. *Journal of Comparative and Physiological Psychology*, 47(6):419–427.

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.

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

Soares, N. (2015). The Value Learning Problem. Technical report, MIRI.