Inferring human values for safe AGI design

Can Eren Sezener

erensezener@gmail.com

Intelligence & goals

Intelligence cannot be defined in the absence of goals.

[Legg and Hutter 2007]

Intelligence & goals

Intelligence cannot be defined in the absence of goals.

[Legg and Hutter 2007]

During this talk, goals $\equiv$ utilities $\equiv$ rewards $\equiv$ values.

The problem

Whatever the architecture of an AGI is, it will likely have an explicit value function.

What an AGI should value should be similar to what humans value.

The problem: How can an AGI learn what humans value?
(aka *the Value Learning Problem* [Soares 2015])

Human values

Humans have complex value systems [Yudkowsky 2011] and it is shown that humans are unable to determine what they value [Muehlhauser and Helm 2012].

Therefore, crafting utility functions for AGI systems that encapsulate human values by hand is not viable.

Learning Human Values

We can attempt to learn a utility function, $U : \mathit{Perceptions} \rightarrow \mathbb{R}$, in a supervised fashion.

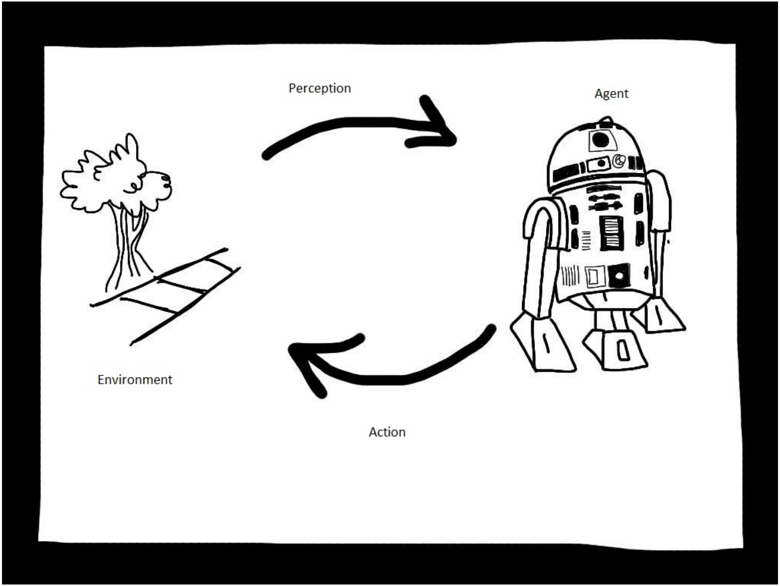
Different ways to do this:

- Ask humans to rate outcomes
- Model humans & ask modeled humans to assign utility values such as in [Hibbard 2012].
- Find a utility indicator (smiles, neuromodulator levels etc.)

However, these have serious shortcomings.

Alternatively, we can directly estimate human values from behavior without requiring revealed preferences.

Agent - Environment



Inverse Reinforcement Learning

Reinforcement Learning: $R \rightarrow \pi^*$
Inverse Reinforcement Learning: $\pi^* \rightarrow R$

Inverse Reinforcement Learning (IRL) is mostly used in robotics.

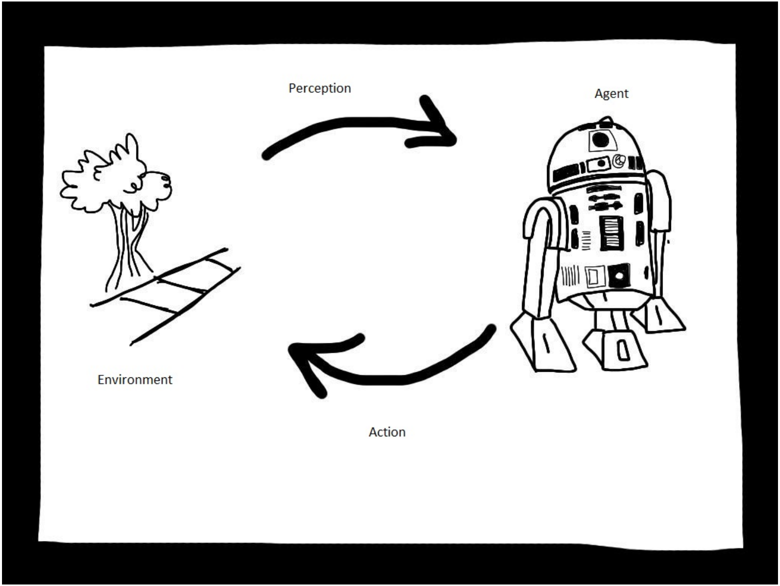
It is recently suggested that IRL might be used to learn human values [Soares 2015].

IRL assumptions

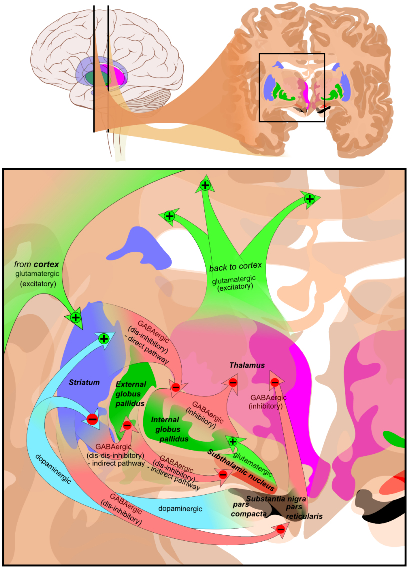
IRL is not feasible for the Value Learning Problem because of its long list assumptions:

- Environment
  - Stationary
  - Fully observable
  - Known (sometimes)
  - Markovian
- Policy
  - Stationary
  - Optimal or near-optimal
- Reward function
  - Stationary

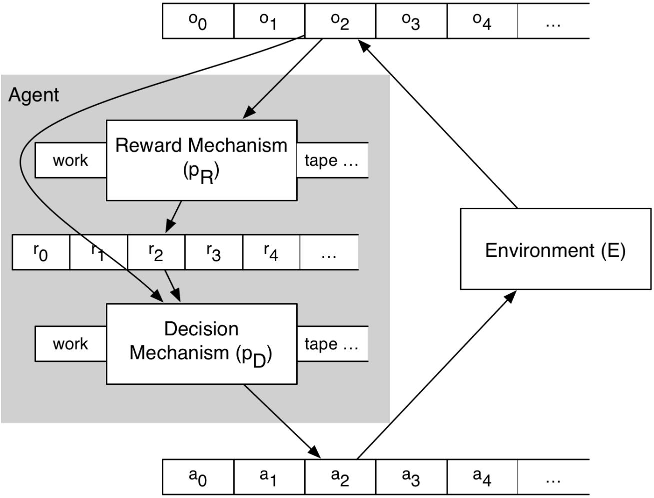
Agent - Environment



The Biological Reward Mechanism



The Universal Reward Inference Framework



The problem: Find the most likely reward mechanism given the action-observation history.
This can be solved with Solomonoff's induction [Solomonoff 1964].

$$M(x) := \sum_{p:U(p)=x*} 2^{-l(p)}$$

is the universal prior where $l(p)$ is the length of the minimal program p, $U(p)$ is the output of a UTM that simulates p, and $x*$ is a string with the prefix x.

Universal IRL

Let's define a joint prior: $m(p_D, p_R) = 2^{-(l(p_D)+l(p_R))}$.
Then, we can solve our problem with

$$m(p_R || a_{1:n}, o_{1:n}) := \sum_{p_D : p_D(p_R(o_{1:n}), o_{1:n}) = a_{1:n}} 2^{-(l(p_R)+l(p_D))}$$

where $a_{1:n} := a_1 a_2 \dots a_n$, $o_{1:n} := o_1 o_2 \dots o_n$, and $p_R(o_{1:n}) = r_1 r_2 \dots r_n$.

Estimating human values

1. Pick N humans
2. Capture all the I/O of those humans
3. Estimate their values
4. Preprocess the values
5. Combine the values
6. Give their values to an AGI

Hibbard, Bill (2012). "Avoiding Unintended AI Behaviors". English. In: *Artificial General Intelligence*. Ed. by Joscha Bach, Ben Goertzel, and Matthew Ikl. Vol. 7716. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 107–116. ISBN: 978-3-642-35505-9. DOI: 10.1007/978-3-642-35506-6_12. URL: http://dx.doi.org/10.1007/978-3-642-35506-6_12.

Legg, Shane and Marcus Hutter (2007). "Universal Intelligence: A Definition of Machine Intelligence". English. In: *Minds and Machines* 17.4, pp. 391–444. ISSN: 0924-6495. DOI: 10.1007/s11023-007-9079-x. URL: http://dx.doi.org/10.1007/s11023-007-9079-x.

References II

Muehlhauser, Luke and Louie Helm (2012). "The Singularity and Machine Ethics". English. In: *Singularity Hypotheses*. Ed. by Amnon H. Eden et al. The Frontiers Collection. Springer Berlin Heidelberg, pp. 101–126. ISBN: 978-3-642-32559-5. DOI: 10.1007/978-3-642-32560-1_6. URL: http://dx.doi.org/10.1007/978-3-642-32560-1_6.

Soares, Nate (2015). *The value learning problem*. Tech. rep. Berkeley, CA: Machine Intelligence Research Institute.

Solomonoff, R.J. (1964). "A formal theory of inductive inference. Part I". In: *Information and Control* 7.1, pp. 1 –22. ISSN: 0019-9958. DOI: http://dx.doi.org/10.1016/S0019-9958(64)90223-2. URL: http://www.sciencedirect.com/science/article/pii/S0019995864902232.

References III

Yudkowsky, Eliezer (2011). "Complex Value Systems in Friendly AI". English. In: *Artificial General Intelligence*. Ed. by Jürgen Schmidhuber, Kristinn R. Thrisson, and Moshe Looks. Vol. 6830. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 388–393. ISBN: 978-3-642-22886-5. DOI: 10.1007/978-3-642-22887-2_48. URL: http://dx.doi.org/10.1007/978-3-642-22887-2_48.

Thank you.