AI Safety

AISFP 2017

September 13, 2017

Abstract

Abstract here.

Contents

1 Cartesian and naturalized agents

1. Consider a robot thinking about a video game: what inputs would cause me to win?

   It thinks about what happens if it puts in inputs A and B; which input would give good outputs? It does something like argmax.

2. Now consider a different world. In this world the robot imagines itself in this world. The game involves robots thinking about things like robots.

In the first setting, the robot thinks: Here's one thing I can do, and what it does to the world. It has copies of the world inside it.

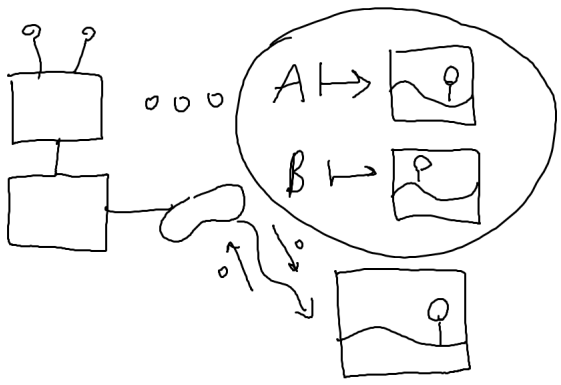

Figure 1: Cartesian agent

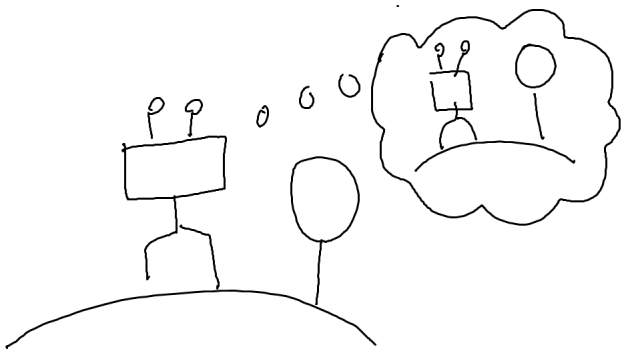

Figure 2: Naturalized agent

In the second setting: You can't fit a whole copy of the world inside the robot's brain because the robot's brain is in the world. There are no "functions." We don't have a function the robot knows or a well-defined function the robot can learn about. There's nothing that feels like a "base" function.

In the second world, there could be other things that reason like the robot.

The first setting is a **Cartesian agent**—"plug into" the world. The second is a **naturalized agent**.

Question: what are naturalized agents like?

(You can introduce some complications in the Cartesian case by requiring the robot to be smaller than the world.)

There could be a thing in the video game that is a copy of the robot: the robot affects its input-output channel but also something else in the video game.

1.1 Discussion

1. Q: In practice don't fully simulate the world/other agents. What about a partial simulation? But some robot which did more can exploit you?

   The mindset is: We're focusing on creating Euclidean geometry, not do the engineering.

2. Only thing that controls is input/output. Contents of agent's head and thought process. You can't split it into input/output channel.

3. Imagine I just thought of the world as neurons firing. But isn't that isomorphic to knowing about myself? The thought itself is an action.

   "You're optimizing the thing that is doing the optimizing."

4. The field of game theory has part of "naturalized agency," making the first picture look more like the second. Other agents are off the same type as you, but you view them as perfect adversaries.

5. Why is this interesting from AI alignment perspective?

Objection: It's better to stick together already existing tools and approaches/bits of confusion (Nash equilibrium, find more things to collide it with) than create something new from scratch.

6. The I/O (Cartesian) model makes it difficult to think about cases when agent is rewriting its own internals. it doesn't help when the agent starts to grow up.

Here's an outline.

1. Cartesian world: First we look at what we understand about the first picture. That took a lot of time but we understand it well now. This gives us something to compare to.

   What parts of what we understand here get messed up in the second picture, and what keeps working?

   Then we do math.

2. Fixed point: Has most useful tools to think about this. Think about how fixed point theory applies to the picture.

3. Philosophy