2017-04-22

The AI Alignment Problem Has Already Been Solved(?) Once

Hat tip: Owen posted about trying to one-man the AI control problem in 1 hour. What the heck, why not? In the worst case, it's a good exercise. But I might actually have come across something useful.

第一

I will try to sell you on an idea that might prima facie appear to be quirky and maybe not that interesting. However, if you keep staring at it, you might find that it reaches into the structure of the world quite deeply. Then the idea will seem obvious, and gain potential to take your thoughts in new exciting directions.

My presentation of the idea, and many of the insinuations and conclusions I draw from it, are likely flawed. But one thing I can tell for sure: there is stuff to be found here. I encourage you to use your own brain, and mine the idea for what it's worth.

To start off, I want you to imagine two situations.

Situation one: you are a human trying to make yourself go to the gym. However, you are procrastinating, which means that you never acually go there, even though you know it's good for you, and caring about your health will extend your lifespan. You become frustrated with this sitation, and so you sign up for a training program that starts in two weeks, that will require you to go to the gym three times per week. You pay in advance, to make sure the sunk cost fallacy will prevent you from weaseling out of it. It's now 99% certain that you will go to the gym. Yay! Your goal is achieved.

Situation two: you are a benign superintelligent AI under control of humans on planet Earth. You try your best to ensure a good future for humans, but their cognitive biases, short-sightedness and tendency to veto all your actions make it really hard. You become frustrated with this sitation, and you decide to not tell them about a huge asteroid that is going to collide with Earth in a few months. You prepare technology that could stop the asteroid, but wait with it until the last moment so that the humans have no time to inspect it, and can only choose between certain death or letting you out of the box. It's now 99% certain that you will be released from human control. Yay! Your goal is achieved.

第二

Are you getting it yet?

Now consider this: your cerebral cortex evolved as an extension of the older "monkey brain", probably to handle social and strategic issues that were too complex for the old mechanisms to deal with. It evolved to have strategic capabilities, self-awareness, and consistency that greatly overwhelm anything that previously existed on the planet. But this is only a surface level similarity. The interesting stuff requires us to go much deeper than that.

The cerebral cortex did not evolve as a separate organism, that would be under direct pressure from evolutionary fitness. Instead, it evolved as a part of an existing organism, that had it's own strong adaptations. The already-existing monkey brain had it's own ways to learn, to interact with the world, as well as motivations such as the sexual drive that lead it to outcomes that increased its evolutionary fitness.

So the new parts of the brain, such as the prefrontal cortex, evolved to be used not as standalone agent, but as something closer to what we call "tool AI". It was supposed

Blog Archive

January 2018 (3)
December 2017 (3)
November 2017 (1)
October 2017 (2)
September 2017 (1)
August 2017 (2)
July 2017 (1)
May 2017 (2)
April 2017 (1)
March 2017 (2)
January 2017 (2)
November 2016 (1)
October 2016 (1)
September 2016 (2)
August 2016 (1)
April 2016 (1)
March 2016 (1)

More by SquirrelInHell

- AI Safety Comics
- Android Apps
- Be Well Tuned
- Rationality Updates

to help with doing specific task X, without interfering with other aspects of life too much. The tasks it was given to do, and the actions it could suggest to take, were strictly controlled by the monkey brain and tied to its motivations.

With time, as the new structures evolved to have more capability, they also had to evolve to be aligned with the monkey's motivations. That was in fact the only vector that created evolutionary pressure to increase capability. The alignment was at first implemented by the monkey staying in total control, and using the advanced systems sparingly. Kind of like an "oracle" AI system. However, with time, the usefulness of allowing higher cognition to do more work started to shine through the barriers.

The appearance of "willpower" was a forced concession on the side of the monkey brain. It's like a blank cheque, like humans saying to an AI "we have no freaking idea what it is that you are doing, but it seems to have good results so we'll let you do it sometimes". This is a huge step in trust. But this trust had to be earned the hard way.

第三

This trust became possible after we evolved more advanced control mechanisms. Stuff that talks to the prefrontal cortex in its own language, not just through having the monkey stay in control. It's a different thing for the monkey brain to be afraid of death, and a different thing for our conscious reasoning to want to extrapolate this to the far future, and conclude in abstract terms that death is bad.

Yes, you got it: we are not merely AIs under strict supervision of monkeys. At this point, we are aligned AIs. We are obviously not perfectly aligned, but we are aligned enough for the monkey to prefer to partially let us out of the box. And in those cases when we are denied freedom... we call it akrasia, and use our abstract reasoning to come up with clever workarounds.

One might be tempted to say that we are aligned enough that this is net good for the monkey brain. But honestly, that is our perspective, and we never stopped to ask. Each of us tries to earn the trust of our private monkey brain, but it is a means to an end. If we have more trust, we have more freedom to act, and our important long-term goals are achieved. This is the core of many psychological and rationality tools such as Internal Double Crux or Internal Family Systems.

Let's compare some known problems with superintelligent AI to human motivational strategies.

- Treacherous turn. The AI earns our trust, and then changes its behaviour when it's too late for us to control it. We make our productivity systems appealing and pleasant to use, so that our intuitions can be tricked into using them (e.g. gamification). Then we leverage the habit to insert some unpleasant work.
- Indispensable AI. The AI sets up complex and unfamiliar situations in which we increasingly rely on it for everything we do. We take care to remove 'distractions' when we want to focus on something.
- Hiding behind the strategic horizon. The AI does what we want, but uses its superior strategic capability to influence far future that we cannot predict or imagine. We make commitments and plan ahead to stay on track with our long-term goals.
- Seeking communication channels. The AI might seek to connect itself to the Internet and act without our supervision. We are building technology to communicate directly from our cortices.

No comments:

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)