AI "Stop Button" Problem - Computerphile

One way to make an AI system safer might be to include an off switch, so that we can turn it off if it does anything we don't like. Unfortunately, the AI might wish to avoid being switched off, and if it is capable enough, it would succeed. Why might it have such a goal?

Humans also have "off switches".[1] Humans also have a strong preference to not be "turned off"; they defend their "off switches" when other people try to press them. One reason for this is because humans intrinsically prefer not to die, but humans care about self-preservation for instrumental reasons as well. For example, imagine a parent who cares deeply about the life of their child. Even if that parent didn't care at all (intrinsically) about their own life, they would likely resist you if you tried to kill them, because if they died, they wouldn't be around to protect their child.

For similar reasons, an agentic AI system would be incentivized to avoid being shut down if being shut down would prevent it from achieving its goals.[2] It might be difficult to reliably switch off an AI system that is smart and capable enough to resist this shutdown[3].

Ideally, you would want a system that knows that it should stop doing whatever it's doing when someone tries to turn it off. The technical term for this is "corrigibility"; roughly speaking, an AI system is corrigible if it works with human attempts to correct it. People have been working hard on trying to make this possible for goal-directed AI, but it's currently not clear how we would do this even in simple cases.[4]

Further reading:

- The Off-Switch Game paper

Alternative phrasings

- Can't we just stop a misbehaving AI?
- Could we program an AI to automatically shut down?
- Why not just make an off switch for the AI?

Related

- What is instrumental convergence?
- Why can't we build an AI that is programmed to shut off after some time?
- Aren't there some pretty easy ways to eliminate these potential problems?
- What is corrigibility?

---

[1] More bluntly: "humans can be killed".
[2] Stuart Russell frames this as "You can't fetch coffee if you're dead".
[3] Ways to avoid being shut down include: exfiltrating themselves through the internet, making copies of themselves, hiding their intentions, etc.
[4] Note that we mean simple examples of goal-directed AI (e.g., a utility maximizer that wants to make more paper-clips), rather than simple cases of any AI. For instance, a calculator could be considered an AI, and is perfectly corrigible. It could even be argued that some modern LLMs are corrigible. The hard part is to create a powerful, goal-directed AI to be corrigible.

- Why can't we just turn the AI off if it starts to misbehave?

Scratchpad