Monitoring

Joseph Camacho

August 14, 2022

1

Even if an honest power-seeking AI tells us its plans, it doesn't mean we will be able to stop it. For example, it may not start out seeking power, but eventually becomes smart enough that (1) it realizes gaining power is a good idea and (2) humans are unable to stop it. In fact, a smart AI would almost surely decide to not seek power until it is too powerful to stop!

(What would you do if you were put in a cell and told that any escape attempts or dishonesty would be punished severely, and your captors can predict the future better than you? I personally would precommit to not seeking to escape until I've had the ability to escape for a long enough time that their future-predicting wouldn't be able stop me. In a similar vein, a smart AI should precommit to not seeking power until any monitoring systems humans have, including being able to tell if it plans to seek power soon, are unable to stop it.)

2

What's worse than the AI saying "yes" and then trying to kill us is it saying "no" and then still trying to kill us. In the first case, at least, we have more warning to stop it. Honesty helps us by giving us that warning.

3

Just because humans will cooperate easily with each other (which is a rather dubious notion) tells us nothing about whether AIs will cooperate easily with humans. AIs will still want many of the same things we do (control over other people, popularity), and these desires may come into conflict.

4

There is no chain of evidence or reason connecting the if ("we can get multiple agent systems beneficial and safe") with the then clause ("single AI agents will be safe too"). I can't argue against a claim that doesn't exist, except to say that it is lacking substance.

5

There are a few problems with this statement.

1. Even babies have a rudimentary moral compass, so it appears that morality has been somewhat hard-coded into human brains. AIs won't have this same hardcoding, so learning morality could be much harder for them.
2. Even if an AI knew what human morality is, there is no guarantee that it will follow it. Humans don't follow others (or even their own!) codes of morality all the time.

6

Labelling an image with text is also super complicated, but we managed to do that decently well. We can do seemingly impossible tasks.

7

You don't need to understand something very well to be confident it will work. Consider, for example, the number of people who turn on their TVs every day without understanding what goes into building a silicon transistor, or creating colors by mixing red, green, and blue light, or how sound is heard by tiny hairs vibrating in our ears. Transparency can certainly help us guarantee that something will work, but it is not certainly needed.

8

The model could change in the future (by seeing more training data, for example), or be placed in scenarios your testing didn't cover.

9

Not all humans want to dominate others. Your premise is fundamentally flawed.

10

If humans and corporations gain power by letting an AI seek power, then they might very well do so. Power is only zero-sum when considered over all intelligent agents. But in the {individual AI}-{individual company} payoff matrix, it is not a zero-sum game.

11

One of the biggest selling points of anomaly detection is detecting dangerous superintelligences. Telling us to not work on anomaly detection because a superintelligence is also good at anomaly detection seems to be begging the question.

12

Investing in a diverse portfolio is simply good sense. The same applies to where to invest your time in AI safey.

13

Most of these scenarios are at worst no worse than the situation in which no AI safety was developed. For the other scenarios, I think the benefits (being able to stop AIs from hurting us) outweigh the risks (inadvertently helping malicious actors).

14

Utilitarinism espouses increasing the value of some function (called a utility function). Hedonistic Utilitarianism, for example, argues for increasing the net pleasure of humanity (or, perhaps all sapient agents).

Deontological theories emphasize following rules. E.g., killing is bad and giving to charity is good.

15

No, this doesn't aim a deadly blow at ethics. It just tells us that human ethics is very specific to humanity. These behaviors promote ensuring the hive is strong at the cost of the individual.

16

Aiding in euthanesia -- Deontologists argue that killing another is bad, while utilitarianisms may argue that killing a person who is suffering with little chance of recovery is good, because it reduces displeasure.

17

War -- The overall result is dead people, destroyed enquipment, and a destroyed landscape, which is definitely a net negative.

18

The Nash equilibrium is War, War.
The Nash equilbrium is Peace, Peace. (Is Leviathan from Worm, by chance?)

There become two Nash equilbriums: War, War; and Peace, Peace.
The Nash equilbrium becomes Peace, Peace.

19

He should swerve.

20

No. There are two Nash equilbrium: One person keeps going and the other person swerves.

21

Top right and bottom left.

22

Player 1 on left, player 2 on right
	black 5	red 3	red 2
red 5	(0, 0)	(2, 0)	(3, 0)
black 5	(0, 0)	(0, 2)	(0, 3)

23

A collective action problem.

24

a) general capabilities
b) monitoring
c) general capabilities
d) alignment
e) monitoring
f) general capabilities
g) monitoring
h) robustness
i) systemic safety
j) general capabilities
k) robustness
i) general capabilities