Why could AI "want" to kill everyone? - YouTube

Artificial intelligence (AI) is a transformative technology, and as such, it comes with risks. People can use AI to do bad things, or the use of AI systems can result in bad outcomes due to human error. The same is true of any very powerful technology.

However, there are challenges that are unique to AI. Unlike other technologies, AI can be given the ability to make autonomous decisions, which introduces the possibility of actions that were never intended by its creators. This challenge of ensuring AI's actions align with human intentions is termed the "alignment problem," and the associated risks are often called "accident risks."[1] [2] The focus on this site is primarily on the risk that an AI might do bad things even if its designers only intend to use it for good.

To give a simplistic, but illustrative example: if an AI is given the goal of eradicating cancer in humans, the most reliable way to do so might be to kill humanity instead of finding a cure.

Complexity of Value: Human values aren't as simple as they may seem

A quick fix for this behavior might look like changing the AI's goal from "eradicate cancer" to "eradicate cancer while preserving human flourishing". But here we run into another roadblock: we don't know how to mathematically define human concepts like "flourishing", and as a result, we can't explicitly program them into the AI.

Orthogonality Thesis: Intelligence and human values need not go together

There might come a point when an AI is intelligent enough to understand exactly what we mean by "preserve human flourishing". However, just because the AI understands what we intend, that doesn't mean it will actually do what we intend, because it won't "care" about following our intent unless we explicitly train it to. Unfortunately, we don't presently know how to train an AI to "care" about following our intentions, and we can't expect it to care by default, because intelligence and morality are "orthogonal" – they can vary separately, as though they were different dimensions at right angles to each other.

The complexity of value and orthogonality theses, taken together, suggest that ensuring an AI's goals match our intentions could be hard. Furthermore, if its goals don't match our intentions, there is reason to think that the most effective actions it could take toward its own goals would be harmful to us.

Instrumental Convergence: Certain bad behaviors are useful for most AI goals

For almost any final goal that an AI could have, there are instrumental goals it could pursue that make it more likely to achieve that goal, such as acquiring resources, staying

[1] "Accident risks" here do not refer to accidental human actions like someone pressing the wrong button; such incidents fall under "misuse risks" – because it is an action taken by a human that caused the problem.
[2] The distinction between misuse and accident risk isn't clear-cut. See structural risks.

operational, and preserving its goals. This could be catastrophic for humanity, because an AI with these instrumental goals may end up using resources humans need to live. Moreover, it may take extreme measures to prevent us from turning it off or altering its goals, because human interference might lead to its final goal not being achieved.

All in all, an AI does not need to be malevolent to do bad things – it's enough for it to be indifferent to what we care about. This makes it crucial to align AI with human values.

Related

- Why would a misaligned superintelligence kill everyone in the world?
- What are accident and misuse risks?
- Is AI safety about systems becoming malevolent or conscious and turning on us?
- What is instrumental convergence?
- What is the orthogonality thesis?

Scratchpad

Murphant's 2024-08 opinion is that

Why would an AI do bad things?
Why might we expect a superintelligence to be hostile by default?
Why might a superintelligent AI be dangerous?
are all the same questions, and the articles should be either merged or differentiated