There are several broad dynamics that seem like plausible contributors to the risk of an AI-caused existential catastrophe.

There are a number of ways that AI could end up behaving dangerously:

- Training processes could produce unintended outcomes, for example a "misaligned mesa-optimizer"
- We could misspecify our goals, producing an AI that pursues goals we don't want
- AIs could be misused by people intentionally trying to cause harm

Additionally, there are features of the world that could make avoiding a disaster harder:

- Insufficient time to solve open technical problems, especially around AI alignment.
- A lack of coordination between the most important actors, like AI labs and national governments.
- The acceleration of progress through cheaper computing hardware, algorithmic progress and increased investment

One could also look at different kinds of dangerous uses AI could be put to, like locking in undesirable values or inventing powerful weapons. Different types of errors could persist in an AI even as its capabilities became highly advanced, like incorrect assumptions about metaethics, decision theory, or metaphilosophy.

A post-AGI world could end up with different broad patterns where human values lose influence, like new competitive pressures or concentration of power.

Related

- What are existential risks (x-risks)?
- What are accident and misuse risks?
- How likely is extinction from superintelligent AI?

Scratchpad

<Siao's 2024-01 draft with updates to the text above>

While we can't predict the full story of how AI will affect the future, there are several broad dynamics that seem like plausible building blocks of an existential catastrophe.

One perspective is to viewFor example, at the time advanced AI is invented, some features of the world that could make avoiding disaster much harder as sources of existential risk. For example, we might end up with: —

- Insufficient time to solve the problem,
- Insufficient coordination between the most important actors

- , Mmuch cheaper computing hardware—could make avoiding a disaster much harder.

Another perspective is to look at different ways a dangerous AI could come about:

- Throughfor example, as an inner optimizationzer,
- or bBecause we accidentally misspecify our values, or through misuse. And one could look at different kinds of dangerous uses it could be put to, like locking in undesirable values or inventing powerful weapons.

Different types of errors could persist in an AI even as its capabilities became highly advanced, like bad assumptions about metaethics, decision theory, or metaphilosophy.

A post-AGI world could end up with different broad patterns where human values lose influence, like new competitive pressures or concentration of power

<end Siao draft>