Fire can be used to cook food or to burn down a village, and nuclear reactions can be used to produce electricity or city-devastating explosions.

Misuse of AI has the potential to cause harm at the individual, societal, and global levels by enabling:

- Authoritarian surveillance
- Terrorism
- Autonomous weapons
- Non-consensual intimate deepfakes
- Misinformation and propaganda

One of the aims of AI governance is to find ways to prevent such harm.

However, misuse is not the only risk from AI. Even when used with the best of intentions, technology can cause accidents: a fire can spread to burn down a city, and a reactor can melt down.

Dangers from advanced AI could be particularly bad when AI can pursue its own goals, since we currently don't know how to specify safe goals and have highly advanced systems follow them. Future AI could be powerful enough to outmaneuver humanity, and accidents resulting from misalignment could threaten human civilization as a whole.

In other words, for advanced AI to go well, we will need to both:

1. Coordinate to prevent bad actors from misusing AI.
2. Ensure, by solving the alignment problem, that powerful AI can be safely used at all, without escaping human control and causing catastrophic harms that no human intended.

Alternative phrasings

- What about bad actors?

Related

- What are accident and misuse risks?
- Isn't the real concern with AI something else?
- Isn't the real concern autonomous weapons?

Scratchpad