AI Ethics and Value Alignment for Nonhuman Animals

AI Ethics and Value Alignment for Nonhuman Animals

Soenke Ziesche

Episode #28: Digital transformation and Ethical use of technology for animals

26 July 2023

The AI value alignment problem

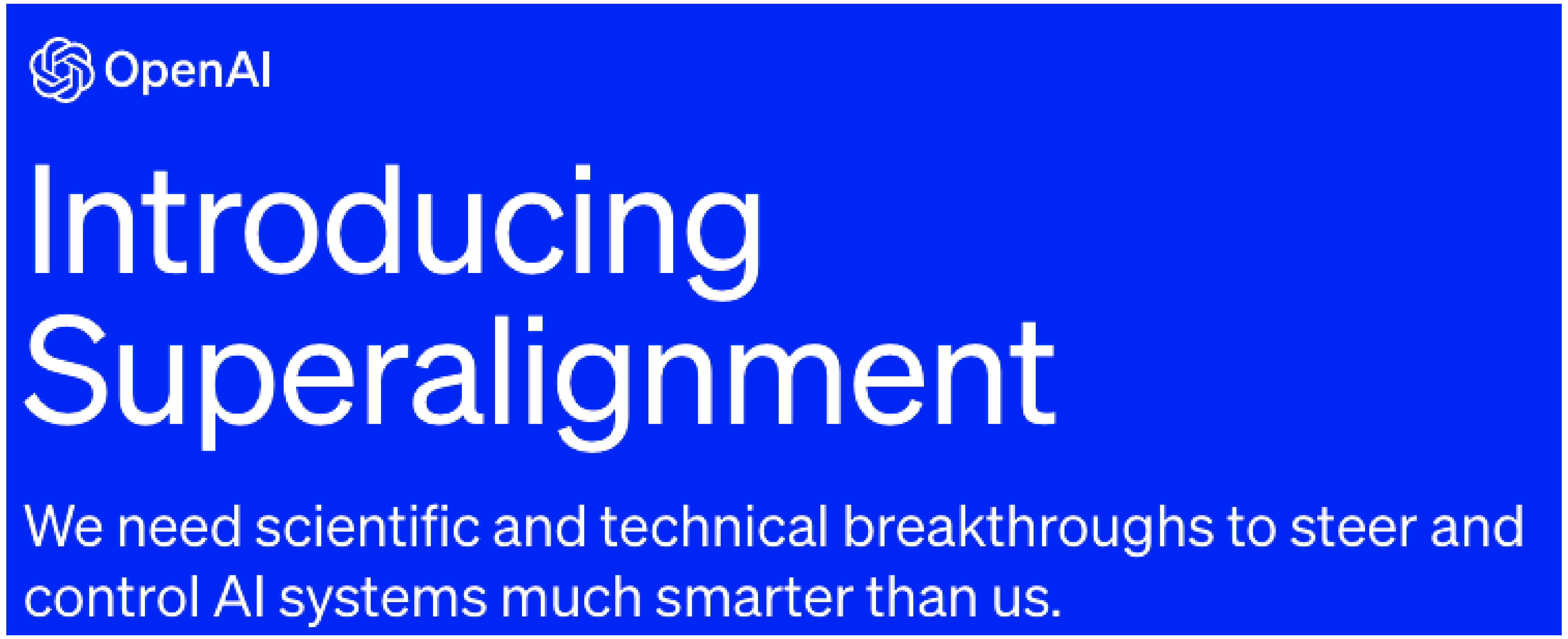
The AI value alignment problem

How can we ensure that AI systems, pursue goals and values which are aligned with **human** goals and values, especially for not yet developed super intelligent Artificial General Intelligence systems?

1) Extract values
2) Aggregate values
3) Instil values into AI systems

E. Yudkowsky:

"The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else."



OpenAI

Introducing Superalignment

We need scientific and technical breakthroughs to steer and control AI systems much smarter than us.

Value extraction for nonhuman animals

Value extraction for nonhuman animals

Challenges

- To actually identify the values and interests of nonhuman animals.
- To explore a large variety of species of nonhuman animals, thus likely a large variety of values and interests.
- To focus on endangered species, assuming that self-preservation is a preference of all nonhuman animals.

Value aggregation for nonhuman animals

Value aggregation for nonhuman animals

Challenges

- To address the fact that values and interests of humans and nonhuman animals are conflicting.
- To address the fact that certain short-term preferences of nonhuman animals are not necessarily good for their long-term health.
- To address the fact that values and interests between species of nonhuman animals are not only different, but also conflicting, e.g., that predators kill and consume prey.

Summary

Summary

- Without value alignment, AI systems would have little incentive to be cooperative or to be altruistic towards nonhuman animals.
- Therefore, AI systems may become existential as well as causing suffering risks for nonhuman animals.
- Also, other existential AI risks, apart from value alignment, such as malicious use or AI race are currently only examined towards safety of humans, but may affect nonhuman animals too.

Proposed scenario

Proposed scenario

Value-aligned AI systems to become custodians for nonhuman animals and for anti-speciesism in general.

- Humans are biased and inconsistent when it comes to the treatment of nonhuman animals since many humans have an interest in their meat, eggs, milk, fur, leather, wool, etc. as well in the habitats of nonhuman animals.

- Humans have incomplete knowledge about how to ensure animal welfare, which includes the largely neglected suffering of wild animals, while AI systems may find solutions in this regard.