If programmed with the wrong motivations, a machine could be malevolent toward humans, and intentionally exterminate our species. More likely, it could be designed with motivations that initially appeared safe (and easy to program) to its designers, but that turn out to be best fulfilled (given sufficient power) by reallocating resources from sustaining human life to other projects. As Yudkowsky writes, “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.”

Since weak AIs with many different motivations could better achieve their goal by faking benevolence until they are powerful, safety testing to avoid this could be very challenging. Alternatively, competitive pressures, both economic and military, might lead AI designers to try to use other methods to control AIs with undesirable motivations. As those AIs became more sophisticated this could eventually lead to one risk too many.

Even a machine successfully designed with superficially benevolent motivations could easily go awry when it discovers implications of its decision criteria unanticipated by its designers. For example, a superintelligence programmed to maximize human happiness might find it easier to rewire human neurology so that humans are happiest when sitting quietly in jars than to build and maintain a utopian world that caters to the complex and nuanced whims of current human neurology.

Related

- How could an intelligence explosion be useful?
- What is an intelligence explosion?

Scratchpad

See also:

- Yudkowsky, Artificial intelligence as a positive and negative factor in global risk
- Chalmers, The Singularity: A Philosophical Analysis