



Related

-  Rob Miles - Why should I care about AI safety?
-  What evidence do experts usually base their timeline predictions on?
- How long until AGI?
- What does “timelines” mean?
- Does AI safety research bring profit?

Scratchpad

Are there “selfish” reasons for caring about AI safety?

Yes.

Positive effects of AI alignment:

- Unlocks the power of AI, depending on how cautious you are forced to be
(the less misalignment you tolerate (or are allowed to tolerate), the more AI models that you want to use or sell or invent, will be banned, weakened or not invented at all, due to safeguards and regulations)
- Controlling AI is closely related to AI safety, and control is useful
- Similarly for interpretability: it is closely related to AI safety, and the better you understand your AI, the better you can put it to use

Negative effects of AI misalignment:

- Misaligned AI is simply bad for business. And as AI systems grow more powerful over time, customers will tolerate even less risk of bad or unpredictable behavior
(yes there is sometimes entertainment value in “bad behavior”, but this is not generally the case)
- In the most catastrophic outcomes, everyone or nearly everyone could suffer and/or die, which includes selfish people

Things to elaborate on:

- Control/interpretability is useful
- What do I mean by selfish? (money, safety, safety of loved ones, research potential, entertainment)

General thoughts:

- Article feels messy or unstructured (scratchpad-y)
- I want to separate x/s-risk from smaller misalignment issues
- “Unlocks the power of AI” needs to be explained better

Alternative phrasings

- Should selfish people care about AI safety?
- Are there “selfish” reasons for caring about AI safety?