In March of 2023, the Future of Life Institute put out an open letter calling for a pause on "giant AI experiments" of "at least six months”. Less than a week later, AI researcher Eliezer Yudkowsky, writing in *Time* magazine, argued that AI labs should "shut it all down."[1] Not building AGI is certainly a live idea on the table.

But this isn’t a simple proposal to implement, because avoiding dangers from unaligned AGI requires that *no one* ever builds unaligned AGI. There are strong competitive pressures to produce more capable AI, and individual companies or labs might worry that if they stop researching AGI, they’ll be overtaken by others who are more willing to push forward. Additionally, AI researchers who make their living researching AI might be (understandably) reluctant to simply stop.

The field of AI governance includes work on solving these kinds of coordination problems.

Related

- Would a slowdown in AI capabilities development decrease existential risk?
- What are the main problems that need to be solved in AI safety?
- Why can’t we just do x?
- What is an alignment tax?

Scratchpad

Siao draft ~2023-08

“***Misjudgment****:* Assessing the consequences of AI deployment may be difficult (as it is now, especially given the nature of AI risk arguments [3]), so some organizations could easily get it wrong—concluding that an AI system is safe or beneficial when it is not.

“Winner-take-all” competition*:* If the first organization(s) to deploy advanced AI is expected to get large gains, while leaving competitors with nothing, competitors would be highly incentivized to cut corners in order to be first [4]—they would have less to lose.

Externalities*:* By default, actors who deploy advanced AI first by cutting corners would stand to receive all of the potential benefits of their deployment, while only incurring a small fraction of the added global risk (especially if they are only concerned with the interests of a small group).

[1] Some people have been even less diplomatic.

Race to the bottom*:* The above dynamics may involve a dangerous feedback loop. If I expect someone to deploy advanced AI unsafely or to misuse it, I am incentivized to cut corners to beat them to it, even if I am completely informed and concerned about all the risks. After all, I may think that my deployment would be less dangerous than theirs. (And that may incentivize them to cut more corners, in a vicious cycle.) [5]

Delayed safety*:* There may be a substantial delay between when some organization knows how to build powerful AI and when some organization knows how to do so safely. After all, such safety delays are common in many industries. Additionally, it may be infeasible to solve AI safety problems before risky AI capabilities are created, since these capabilities may provide testbeds and tools that are critical for solving safety problems.

> (This delay may be the period of especially high risk; soon after this delay ends, risks from unsafe AI may be greatly reduced, because incentives to deploy it may be lower and safe AI may increase humanity's resilience.)

Rapid diffusion of AI capabilities*:* Soon after some actor becomes capable of deploying unsafe AI, many other actors may also gain that capability. After all, recent AI advances have diffused quickly (including internationally) [7], information security weaknesses could cause AI advances to diffuse even faster, the number of actors explicitly aiming to develop general AI has been increasing, and that trend may accelerate when general AI is seen as being more within reach.

"

(from an alternately phrased question)

We could, but it seems unlikely. Each advance in capabilities which brings us closer to an intelligence explosion also brings in profits for whoever develops it (e.g. smarter digital personal assistants, more ability to automate cognitive tasks, better recommendation algorithms for social media). The incentives are the basic problem here. Each individual actor (nation or corporation) fears that if they stop, they'll just get overtaken by others who are more reckless.