Machine Intelligence Research Institute
548 Market Street PMB 46597
San Francisco, CA, 94104

MIRI Briefing on Extinction Risk from AI

September 2024

I. The development of artificial superintelligence poses an imminent risk of human extinction.

"Artificial superintelligence" (ASI) refers to AI that can substantially surpass humans in all strategically relevant activities (economic, scientific, military, etc.).

The timeline to ASI is uncertain, but probably not long. On the present trajectory, MIRI would be uncomfortable ruling out the possibility that ASI is developed in the next year or two, and we'd be surprised if it was still several decades away.[1]

AI labs are aggressively scaling up systems they don't understand. The Deep Learning techniques behind the rapid AI progress of the last few years create massive neural networks automatically. The resulting models are vast human-unreadable tangles of machine-written operations more "grown" than designed.[2] Labs basically discovered a "cheat": Engineers can't tell you why a modern AI makes a given choice, but have nevertheless released increasingly capable systems year after year.[3]

Sufficiently intelligent AIs will likely develop persistent goals of their own. Humans have wants, and make long-term plans, for reasons that we expect to also apply to sufficiently smart mechanically-grown AIs. (The computer science of this prediction does not fit into a paragraph; inquire further if interested.) We are only barely starting to see this phenomenon in today's AIs, which require a long training process to hammer them into the apparently-obedient form the public is allowed to see.[4]

We expect the the ASI's goals to be hollow and lifeless in the end. Imbuing a superhumanly intelligent AI with a deep, persistent care for worthwhile objectives is much more difficult than training it to answer the right way on an ethics test.[5] Having spent two decades on the serious version of this problem, our informed view is that the field is nowhere near a solution.[6]

ASI that doesn't value us will end us. Unless it has worthwhile goals,[7] ASI will put our planet to uses incompatible with our continued survival, just as we fail to concern ourselves with the crabgrass growing on the site of a planned parking lot.[8] No malice, resentment, or misunderstanding is needed to precipitate our extinction.[9]

II. Human survival likely depends on delaying the creation of ASI as soon as we can for as long as necessary.

A "wait and see" approach to ASI is probably not survivable. A superintelligent adversary will not reveal its full capabilities and telegraph its intentions.[10] It will not offer a fair fight. It will make itself indispensable[11] or undetectable until it can strike decisively and/or seize an unassailable strategic position.[12]

MIRI doesn't see any viable quick fixes or workarounds to misaligned ASI. OpenAI admits that today's most important methods of steering AI won't scale to the superhuman regime.[13] Attempts to restrain[14] or deceive[15] a superior intelligence are prone to fail for reasons both foreseeable and unforeseeable.[16] Our own theoretical work suggests that plans to align ASI using unaligned AIs are similarly unsound.[17] We also don't think a well-funded crash program to solve alignment would be able to correctly identify solutions that won't kill us.[18] Our current view is that a safe way forward will likely require ASI to be delayed for a long time.[19]

Delaying ASI requires an effective worldwide ban on its development, and tight control over the factors of its production. This is a large ask,[20] but domestic oversight, mirrored by a few close allies, will not suffice. This is not a case where we just need the "right" people to build it before the "wrong" people do, as ASI is not a national weapon; it is a global suicide bomb.[21] If anyone builds it, everyone dies.

To preserve the option of shutting down ASI development if or when the will is found, MIRI advocates promptly *building the off-switch.*[22] The "off-switch" refers to the systems and infrastructure needed for the eventual enactment of a ban.[23] It starts with identifying the relevant parties, tracking the relevant hardware, and requiring that advanced AI work take place within a limited number of monitored and secured locations. It extends to building out the protocols, plans, and chain of command to be followed in the event of a shutdown decision. As the off-switch could also provide resilience to more limited AI mishaps, we hope it will find broader near-term support than a full ban.[24]

An off-switch can only prevent our extinction from ASI if it has sufficient reach and is actually used to shut down development in time.[25] If humanity is to survive this dangerous period, it will have to stop treating AI as a domain for international rivalry and demonstrate a collective resolve equal to the threat.

Endnotes

intelligence.org/briefing