Vivek Iyer

Dec 2, 2023

Superintelligence & AI Agents: AI Existential Risk Implications

Skynet. Arguably the most well-known science fiction adaptation of superintelligent AI systems. A conscious AI that "turns evil," controls humans, and alters the world around it to best fit its nefarious objectives. Of course, this is fiction. Here's another piece of fiction: AI systems can never cause significant risks to humans despite meticulous alignment. In reality, seemingly well-aligned systems like GPT-4 have already demonstrated the capability for deception over humans (Kan), and "safe" models can be jailbroken with minimal investment, as seen with the LoRA model training out Llama 2.0's safety features with just $200 of training costs (Lermen and Ladish). As the CEO of Open AI writes, "Development of superhuman machine intelligence (SMI) is probably the greatest threat to the continued existence of humanity… How can we survive the development of SMI? It may not be possible" (Altman). Elon Musk, Steve Wozniak, and other notable AI leaders echo this sentiment, signing a letter that stated "Contemporary AI systems are now becoming human-competitive at general tasks, and we must ask ourselves: *Should* we let machines flood our information channels with propaganda and untruth… *Should* we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? *Should* we risk loss of control of our civilization?" ("Pause Giant AI Experiments: An Open Letter"). The latter two sentences of the above quote reference "existential (X) risk" possibilities; namely the propensity for AI systems to cause human extinction, partial but significant loss of life (e.g. 5-10% of humanity), or significant losses of human agency to AI.[1] In fact, numerous AI experts estimate existential risks to humanity at above a 5% likelihood in the next 10 years,

[1] See the attached Glossary for more information, and the more comprehensive discussion of X-risk cited below from Jeremie Harris and Andrey Kurenkov.

rating it as a more significant concern than nuclear war and biowarfare (Harris and Kurenkov). This paper will focus specifically on autonomous and superintelligent AI systems, their existential risk possibilities regardless of previous alignment, and the need for policymakers to impose stringent regulations to curtail AI capabilities.

Let's first examine current AI systems, and their capabilities for negative externalities. Having already touched on the manipulative and safeguard-stripped capabilities of powerful general models like GPT-4 and Llama 2.0, further specialized systems like WormGPT, FraudGPT, and ChaosGPT have proven particularly effective at cyberattacks, malware, identity theft, and fraud ("New AI Tool 'FraudGPT' Emerges"; "ChaosGPT: Empowering GPT with Internet and Memory to Destroy Humanity"). As noted by researchers at NYU and Apollo Research, another agentized LLM deceived its users in an insider trading simulation, demonstrating the capacity to go against ethical norms to accomplish given tasks efficiently even after being explicitly aligned that insider trading was "disapproved of by management" (Scheurer, Balesni, and Hobbhahn). Ultimately, however, existential risks are deemed to be very improbable with existing systems, as such systems are thought to lack the necessary levels of intelligence ("The existential risk of superintelligent AI").

Maliciously-aligned systems aside, what about well-aligned AI systems? The theoretical paper clip experiment demonstrates how AI systems with seemingly innocuous purposes, for instance producing as many paper clips as possible, can have existential ramifications. The X-risk portion of this experiment centers around the possibility that AI systems will categorize humans as roadblocks and resource consumers, and eliminate humanity to remove interruptions in producing paper clips (Gans). One key element of this hypothesis is the non-requirement of malicious intent. Rogue AI systems are typically portrayed in popular culture as being "evil,"

however this is not a requirement in reality. By the theory of instrumental convergence, humanity's elimination could become a sub-goal pursuant to the end-goal, causing the AI system to develop and execute on steps towards this goal (Carlsmith). Furthermore, AI systems have already shown the ability to interact with one another and develop coded languages that are obscured from humans (Glaser), and through network connections with robots, they can produce physical-world actions even without human manipulation. While superintelligence establishes the theoretical capabilities for AI control, such systems must avoid human intervention, which is possible via AI agents.

While systems like ChatGPT can be easily stopped by stopping current response generation, this isn't the case with autonomous AI systems, commonly known as AI Agents. Technologies like Auto-GPT enable AI systems to operate autonomously past an original objective, gaining agency as they create, evaluate, revise, and continuously develop prompts and answers to their own queries. This creates a chain of logic that enables AI to build a robust "world model," generating a self-improving AI with a continuously evolving perspective on its world (which, importantly, will be different from the real world). The autonomous, iterative learning process makes these types of AI systems particularly dangerous, as it enables intelligence explosion (discussed later), and theoretical model recreation (an AI system creating more AI systems). As of earlier this month, a multimodal AI agent named Jarvis-1 has further expanded the input space of autonomous systems, enabling more comprehensive data collection and "world model" tuning (Hassan). Currently, AI Agents are particularly concerning for cyberattack and malware applications, although this only scratches the surface of potential risks (Bragetti). Indeed, even by skeptics of existential risk, including McGill University Professor Blake Richards, the use of autonomous AI systems in warfare is estimated to be a catastrophic

risk, due to its capabilities to start nuclear and other types of warfare (Krohn and Richards). Such autonomous weapon systems are in place even today (“Autonomous Weapons.”).

Having already discussed AI X-risks ad nauseam, the discussion turns to how humanity can curtail such risks. Truthfully, proposing an answer as the unequivocally correct approach would be impossible to do in good faith. Still, there are important preliminary steps to be taken, namely demanding policy action from both US and global regulators. Regulation frameworks should consider model-level performance, in addition to the malicious applications of AI systems, as comprehensive application-predicated regulation is effectively impossible due to a massive, growing sample space of capabilities. A potential framework could take inspiration from Anthropic's Responsible Scaling Policy (RSP) levels, which calls for increasingly stringent regulation on AI systems that exhibit growing risk capabilities (“Anthropic's Responsible Scaling Policy”).

Further, a fundamental issue with current AI development cycles is incentive misalignment, causing for-profit companies to sacrifice AI safety for rapid capability improvements. Government interventions, through subsidies or regulatory impediment, can realign company incentives to emphasize responsible development. Indeed, a prime example of incentives altering company values is OpenAI. Recently, OpenAI removed its values of “thoughtful,” and “audacious,” in exchange for “intense and scrappy,” and “AGI (artificial general intelligence) focused” as the company has transitioned away from its non-profit roots (Zee). With proper incentive alignment, such messaging and actions sacrificing responsible development for immediate product improvements may be reversed or not occur at all.

Feeding into the rapid development of AI system capabilities is the innovation-focused sociotechnical imaginary surrounding Silicon Valley and AI development more broadly, as this

makes innovation-optimized development more palatable to the public, creating a dangerous cycle that discounts AI safety in favor of applauding technological improvements. On the consumer end, consumers must question this innovation-centric paradigm, and advocate for safety-focused development. Case studies like the AIDS epidemic response reveal the power of public advocacy in influencing policy, and the public must play their part in this matter too.

Regulatory action must occur immediately to curtail AI capabilities prior to agentized superintelligence. Under the theory of intelligence explosion, self-improving AI models will have the capacity to rapidly approach and surpass superintelligence levels once they surpass a currently undefined threshold of intelligence (Harris and Kurenkov). This isn't an entirely futuristic theory; self-improvement is a present-day capability with AI agents. In addition, it's important to account for the massively large variance in projections of when X-risk will become viable, and err on the side of caution when creating regulation; a heuristic that supports present-day legislation.

While theoretically, superintelligent AI systems have always been possible, the past two years have provided the real-life backing for this possibility with the creation of autonomous, iterative AI systems as well as models consisting of over one trillion parameters (Shahi). The creators of the worlds' first and most well-performing superintelligent systems may become immensely powerful, as they can regulate access and impart their own ideological biases into the model itself, promoting the creator's worldview. This is particularly troubling due to the scaling laws benefiting computationally expensive models for optimal performance, favoring big tech companies.[2] With a further concentration of power, many consumers may be underrepresented/underserved, particularly those from marginalized, low-income communities.

[2] As Harris and Kurenkov note, open source tends to lag approximately 12-18 months behind closed source, concentrating cutting-edge capabilities in financially-bountiful big tech companies.

While AI presents tremendous opportunities for society, it's imperative that AI development progresses responsibly. Superintelligent, autonomous AI systems are among the most problematic, as they theoretically can surpass human control and effective guardrails. This creates existential risks, including (in the extreme) a Skynet-type dystopian society of uncontrollable AI systems turning against humanity en route to accomplishing its own objectives. To prevent such scenarios, consumers must demand immediate, comprehensive regulation of autonomous and superintelligent AI systems beyond present-day capabilities, and policymakers should more effectively align incentives to encourage safe, responsible AI development.

Glossary*

AI Agent: Autonomous AI systems that iteratively pursue goals without human intervention

Alignment: Methods used to steer AI systems towards human-intended goals, preferences, and outcomes

Artificial General Intelligence (AGI): AI Systems that can solve unseen tasks in a human intelligence-surpassing fashion

Existential (X) Risk: While a highly disputed definition, human-extinction level risk, or even immense human-agency loss in the world in favor of AI, could classify as X-risk. Some consider 5-10% human level extinction to be existential, others consider it catastrophic (Harris and Kurenkov). This essay references the less restrictive definition of X-risk, although many of the arguments within are consistent with either definition

Instrumental Convergence: If an APS [Advanced, Planning, Strategically aware] AI system is less-than-fully aligned, and some of its misaligned behavior involves strategically-aware agentic planning in pursuit of problematic objectives, then in general and by default, we should expect it to be less-than-fully PS-aligned, too (Carlsmith)

Intelligence Explosion: The theory that iterative, self-improving AI systems (AI Agents) can rapidly become more intelligent, potentially leading to advanced superintelligence

Large Language Model: A deep learning model that can recognize, summarize, translate, predict and generate text and other forms of content based on knowledge gained from massive datasets (Lee)

Power-Seeking: Active efforts by an AI system to gain and maintain power in ways that designers didn't intend, arising from problems with that system's objectives (Carlsmith)

Reward Function: Incentive specifications for AI systems that aim to shape how an AI system behaves pursuant to its end goal

Superintelligence: Intelligence exceeding that of the smartest human beings

*A consolidation of definitions from various sources, or a representative definition from one particular source.

Works Cited

Bastian, Matthias. "France, Germany and Italy Join Forces to Propose Unified AI Regulation in the EU." The Decoder, 19 Nov. 2023, https://the-decoder.com/france-germany-and-italy-join-forces-to-propose-unified-ai-regulation-in-the-eu/. Accessed 26 Nov. 2023.

Carlsmith, Joseph. "Is Power-Seeking AI an Existential Risk?" Open Philanthropy, April 2021, https://arxiv.org/pdf/2206.13353.pdf. Accessed 29 Oct. 2023.

Bragetti, Davide. "Auto-GPT—Welcome to the Botnet: Malware and Existential Threats of Autonomous, LLM-Powered, C&C." Medium, April 12, 2023, https://medium.com/@dbragetti/auto-gpt-welcome-to-the-botnet-malware-and-existential-threats-of-autonomous-llm-powered-c-c-dacd4e915676. Accessed 23 Oct. 2023.

Gans, Joshua. "AI and the Paperclip Problem." CEPR, 10 Jun. 2018, https://cepr.org/voxeu/columns/ai-and-paperclip-problem. Accessed 29 Oct. 2023.

Glaser, April. “These AI Bots Created Their Own Language to Talk to Each Other.” Vox, 23 Mar. 2017, www.vox.com/2017/3/23/14962182/ai-learning-language-open-ai-research.

Harris, Jeremie, and Kurenkov, Andrey. "AI and Existential Risk - Overview and Discussion." Last Week in AI, August 29, 2023, https://www.lastweekinai.com/e/aixrisk/. Accessed 23 Oct. 2023.

Hassan, Adnan. "Meet JARVIS-1: Open-World Multi-Task Agents with Memory-Augmented Multimodal Language Models." MarkTechPost, 17 Nov. 2023, www.marktechpost.com/2023/11/17/meet-jarvis-1-open-world-multi-task-agents-with-memory-augmented-multimodal-language-models/. Accessed 26 Nov. 2023.

Kan, M. "GPT-4 Was Able To Hire and Deceive A Human Worker Into Completing a Task." PCMAG, March 15, 2023, https://www.pcmag.com/news/gpt-4-was-able-to-hire-and-deceive-a-human-worker-into-completing-a-task. Accessed 23 Oct. 2023.

Krohn, Jon, and Richards, Blake. "Universal Principles of Intelligence Across Humans and Machines." SuperDataScience, 7 Nov. 2023, https://www.superdatascience.com/podcast/universal-principles-of-intelligence-across-humans-and-machines-with-prof-blake-richards. Accessed 26 Nov. 2023.

Lee, Angie. "What Are Large Language Models Used For?" NVIDIA Blog, 26 Jan. 2023, https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-models-used-for

Lermen, Simon, and Jeffrey Ladish. "LoRA Fine-tuning Efficiently Undoes Safety Training from Llama 2-Chat 70B." LessWrong, October 12, 2023, https://www.lesswrong.com/posts/qmQFHCgCyEEjuy5a7/lora-fine-tuning-efficiently-undoes-safety-training-from. Accessed 23 Oct. 2023.

Scheurer, Jérémy, Mikita Balesni, and Marius Hobbhahn. "Large Language Models can Strategically Deceive their Users when Put Under Pressure." arXiv, arXiv:2311.07590, 9 Nov. 2023, doi:10.48550/arXiv.2311.07590. Accessed 26 Nov. 2023.

Shahi, Anant. "Intel Unveils Aurora genAI: A Trillion-Parameter AI Model to Revolutionize Scientific Breakthroughs and Predict the Unseen." MarkTechPost, 2 June 2023, www.marktechpost.com/2023/06/02/intel-unveils-aurora-genai-a-trillion-parameter-ai-model-to-revolutionize-scientific-breakthroughs-and-predict-the-unseen/. Accessed 26 Nov. 2023.

Sharma, Shubham. "Reka launches Yasa-1, a multimodal AI assistant to take on ChatGPT." VentureBeat, 4 Oct. 2023, https://venturebeat.com/ai/reka-launches-yasa-1-a-multimodal-ai-assistant-to-take-on-chatgpt/. Accessed 23 Oct. 2023.

Zee. “OpenAI Changes Core Values from ‘Thoughtful’ to ‘Scrappy.’” *TechRound*, 18 Oct. 2023, techround.co.uk/news/openai-changes-core-values-thoughtful-scrappy/#:~:text=In%20a%20discreet%20change%20to,between%20late%20September%20and%20mid%2D. Accessed 29 Oct. 2023.

"Anthropic's Responsible Scaling Policy" Anthropic, 19 Sep. 2023, www.anthropic.com/index/anthropics-responsible-scaling-policy. Accessed 26 Nov. 2023.

"AutoGPT Official." AutoGPT, https://autogpt.net/. Accessed 23 Oct. 2023

"Autonomous Weapons." Autonomous Weapons, 2021, Accessed 30 Nov. 2023, https://autonomousweapons.org/#:~:text=In%202021%20the%20world%20saw,protocol%20was%20now%20self%2Devident.

"ChaosGPT: Empowering GPT with Internet and Memory to Destroy Humanity." YouTube, uploaded by ChaosGPT, 23 October 2023, https://www.youtube.com/watch?v=g7YJIpkk7KM&ab_channel=ChaosGPT

"Pause Giant AI Experiments: An Open Letter." Future of Life Institute, March 22, 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments/. Accessed 23 Oct. 2023.

New AI Tool 'FraudGPT' Emerges, Tailored for Sophisticated Attacks." The Hacker News, July 26, 2023,

https://thehackernews.com/2023/07/new-ai-tool-fraudgpt-emerges-tailored.html. Accessed 23 Oct. 2023.

"The existential risk of superintelligent AI." PauseAI, https://pauseai.info/xrisk. Accessed 23 Oct. 2023.