ITIF | INFORMATION TECHNOLOGY & INNOVATION FOUNDATION

Briefing By:

Hodan Omaar
Senior Policy Analyst
Information Technology and Innovation Foundation

To the:

U.S. Senate AI Insight Forum

Hearing on:

"Risk, Alignment, & Guarding Against Doomsday Scenarios"

December 6, 2023
Kennedy Caucus Room, Russell Senate Office Building
Washington, DC

INTRODUCTION

Leader Schumer, Senator Heinrich, Senator Rounds, Senator Young, and distinguished members of the Senate, thank you for the opportunity to discuss how to safeguard artificial intelligence (AI) to minimize risks and ensure it aligns with its designers' goals. I am Hodan Omaar, senior policy analyst of the Information Technology and Innovation Foundation (ITIF), a technology policy think tank.

FEARS ABOUT GOD-LIKE AI

When it comes to AI, most doomsday scenarios predicting catastrophic outcomes stem from the development of what tech entrepreneur and investor Ian Hogarth dubbed "God-like AI" in a now-viral *Financial Times* article.[1] Talking about artificial general intelligence (AGI) or "superintelligence," Hogarth asserted "A three-letter acronym doesn't capture the enormity of what AGI would represent, so I will refer to it as what is: God-like AI. A superintelligent computer that learns and develops autonomously, that understands its environment without the need for supervision and that can transform the world around it."[2]

There are three broad ways developing God-like AI could hypothetically result in existential harms, which for the purposes of this testimony, means those harms that would annihilate humanity or permanently and drastically curtail its potential:

> **Accidents:** Those creating God-like AI systems could unwittingly develop systems that display unintended and harmful behavior that results in existential or catastrophic harm to human civilization. For example, advanced AI systems that do not "align" with human values (commonly referred to as the alignment problem) may launch (or refuse to launch) military weapons systems.
>
> **Misuse**: Malicious actors, such as a rogue state or terrorist organization, could use a God-like AI system to intentionally cause harm. For example, a malicious actor could use advanced AI capabilities to exploit vulnerabilities in large language models to make them release information on how to design new pathogens that cause mass death.
>
> **Structural disruptions**: God-like AI systems could destabilize the broader environment by creating "structural risks" in harmful ways that do not fall into the accident-misuse dichotomy.[3] For example, AI systems that identify or assess the retaliatory capabilities of an adversarial nation could disturb the equilibrium of mutual assured destruction and drastically increase the risk of a nuclear war.

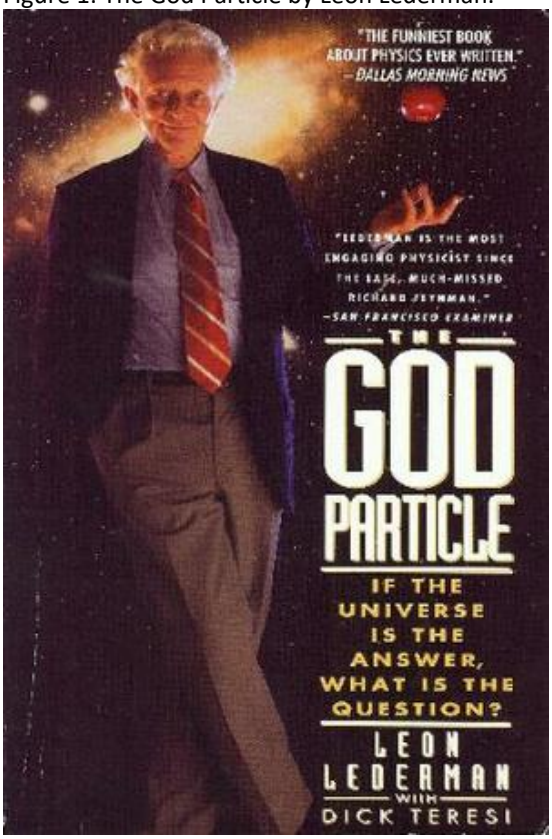
CLAIMS THAT AI DEVELOPERS ARE PLAYING GOD ARE MISLEADING

Critics have often accused scientists, such as those involved in genetic engineering and artificial insemination, of "playing God" to suggest that humans should not tinker with the natural order. Invoking emotive religious metaphors to describe scientific advances in this way is more about conveying a narrative than it is about conveying objective facts. Rather than a reflection of genuine religious beliefs, it uses religion to frame the issue in a deliberately polarizing way. While deemphasizing the advantages that pursuing "God-like" advancements could bring, the use of the term in this context suggests that unlike an omniscient and infallible deity, scientists are prone to errors and limited in their knowledge. It signals scientific hubris will lead to apocalyptic results without their intervention. Policymakers should remember such grandiose terminological and analogical framings are not new. With each technological breakthrough there are often claims that humanity has acquired a capability beyond its control that could ultimately destroy it.

Indeed, new concerns about researchers playing God with AI draw close parallels to concerns that physicists were playing God with particle accelerators, especially the most famous particle accelerator of all, CERN's

Large Hadron Collider (LHC). In 2012, four years after the machine was first switched on, experiments using the LHC helped discover a subatomic particle that widely became known as the "God particle"—a term coined in a 1993 popular science book called *The God Particle: If the Universe is the Answer What is the Question?* (Figure 1).[4] The unfortunate nomenclature made its way into novels of the same name, films, television shows, and shows up in nearly every article on the Higgs Boson, the subatomic particle's proper name.

Figure 1: The God Particle by Leon Lederman.[5]



And much like God-like AI, the LHC "God machine" as it was nicknamed, raised concerns of existential risk. The biggest concern, exacerbated by media articles, was that the particle accelerator might unwittingly generate a black hole that could destroy Earth. And experts indulged the fantasy. A 2001 paper by physicists Stephen Giddings and Scott Thomas noted that if space has more than three dimensions as string theory allows for, then "future hadron colliders such as the Large Hadron Collider will be black hole factories."[6] And a Nobel-prize winning nuclear physicist wrote a 2000 paper titled, "Might a laboratory experiment destroy planet earth?"[7] The concern of black holes destroying Earth was so persistent and oft-raised that CERN still addresses these fears on its website's frequently asked questions (FAQ) page.[8] Black holes were not the only fear; in 2017, satirical website *The Onion* published an article titled "CERN Researchers Apologize For Destruction Of 5 Parallel Universes In Recent Experiment" thumbing its nose at claims the LHC might destroy parallel universes.[9]

The LHC did not create a perilous black hole, and it would have been misguided to attempt to halt physics research to avoid the possibility that sometime in the future a researcher might cause a global disaster. But the absence of catastrophe does not mean that policymakers should ignore potential risks or that responsible oversight of science is unnecessary. Given that some experts are warning of the risks of AGI even while other

experts dismiss them, and since policymakers are unable to know which group is right about the future, it seems policymakers will need to sift through the noise and contend with the potential risks of AGI.

HOW POLICYMAKERS SHOULD CONTEND WITH EXISTENTIAL AI RISKS

There are several proposals for new AI institutions to manage the risks from AI. Some seek to set norms or build political consensus like the G7 Hiroshima AI Process or the UN's AI Advisory Group; some seek to coordinate policy or regulation like a CERN for AI or an International Atomic Energy Agency (IAEA) model for AI; some seek to facilitate standards for AI safety like the newly established UK and U.S. AI Standards Institute; and some seek to build scientific consensus like an Intergovernmental Panel on Climate Change (IPCC) model for AI.[10]

To contend with existential risks, one of the most necessary functions at this stage is to better understand the threat vectors. As one 2020 survey on existential risk scenarios points out, while the original risk scenarios Nick Bostrom, a philosopher at the University of Oxford, and Eliezer Yudkowsky, an AI researcher, put forth have been criticized, the alternative risk scenarios proposed have been largely informal and ""The result is that it is no longer clear which AI risk scenarios experts find most plausible."[11]

The IPCC model for AI is one proposal toward this end but it has challenges. In a *Financial Times* article in November 2023, former Google chief executive Eric Schmidt and cofounder of DeepMind Mustafa Suleyman proposed policymakers should create a new consensus-building institution called the International Panel on AI Safety (IPAIS) modeled on IPCC. They argue, "Before we charge head first into over-regulating we must first address lawmakers' basic lack of understanding about what AI is, how fast it is developing and where the most significant risks lie."[12] A few weeks after the article was published, British Prime Minister Rishi Sunak announced his plan to establish an IPCC for AI at the UK's AI Safety Summit.[13] But critics of this model rightfully point out that an IPAIS would face a much more difficult research environment than the IPCC because unlike the scientifically well-understood mechanisms that underpin climate change, research on AI safety and alignment is still nascent.[14] It might therefore simply end up finding a brittle consensus in a thin research environment. Furthermore, while the authors envision an IPAIS staffed and led by computer scientists and researchers rather than political appointees or diplomats, many have pointed out the potential for the body's research findings to become politicized as they argue the IPCC has.[15]

However, one model that has not yet been invoked but that could be promising to explore, is one based on the Search for Extraterrestrial Intelligence (SETI). The SETI Institute is a nonprofit research organization that was started by a government grant, and unlike an IPCC that would attempt to find scientific consensus, a Search for Artificial General Intelligence (SAGI) Institute would focus on developing benchmarks to determine whether an AI system has reached AGI; identifying when, if ever, anyone develops AGI; and preparing a response plan if that point is ever reached. The SAGI Institute might also serve as a hub for interdisciplinary research, bringing together computer scientists, neuroscientists, ethicists, and other relevant experts to collectively identify the challenges and opportunities associated with AGI.

Importantly, this would differ from institutes that are primarily focused on AI safety, such as the AI Safety Project proposed by DeepMind researcher Lewis Ho and others in a 2023 paper.[16] An AI Safety Project would be an institution with significant compute, engineering capacity, and access to models with the goal of accelerating technical AI safety research. A SAGI model would differ in that it would focus exclusively on the question of detecting signs of AGI and alerting the international community if this event ever occurs, much like SETI has as its mission to alert the world of signs of extraterrestrial life. For example, it might develop a consensus on a definition of AGI (which might include different levels of intelligence), create tests for signs of AGI, and establish a post-detection protocol for how and who to notify if AGI is developed. AI companies could voluntarily participate, and any whistleblowers could also report information. Similar to

SETI's mission, a SAGI would provide a unified forum for those seeking to identify the emergence of AGI, if it is developed. Notably, a SAGI model does not necessarily presuppose AGI is imminent but prepares for the possibility that it could occur.

CONCLUSION

Policymakers should remain clear-eyed in the face of grandiose, uncertain claims. There should always be space to engage with constructive criticisms of specific technological advancements and scientific procedures, as well as potential risk scenarios, but dogmatic narratives should not drown out pragmatic discussions. Cloaking narratives about existential risks of AI in apocalyptic religious language only serves to prevent good-faith efforts to address potential risk scenarios.

REFERENCES

[1] Ian Hogarth, "We must slow down the race to God-like A,I" Financial Times, April 13 2023, https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2.

[2] Ibid.

[3] Remco Zwetsloot and Allan Dafoe, "Thinking About Risks From AI: Accidents, Misuse and Structure," February 11, 2019, *Lawfare blog*, https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure.

[4] Leon Lederman, "The God Particle: If the Universe is the Answer What is the Question?" (Houghton Mifflin, Boston, 1993).

[5] Ibid.

[6] Steven B. Giddings and Scott Thomas, "High Energy Colliders as Black Hole Factories: The End of Short Distance Physics," *American Physical Society*, February 2002, https://doi.org/10.1103/PhysRevD.65.056010.

[7] Francesco Calogero, "Might a laboratory experiment destroy planet earth?" *Interdisciplinary Science Reviews*, July 2013, https://doi.org/10.1179/030801800679224.

[8] "Will CERN generate a black hole?" CERN FAQ page website, accessed November 28, 2023, https://home.cern/resources/faqs/will-cern-generate-black-hole.

[9] "CERN Researchers Apologize For Destruction Of 5 Parallel Universes In Recent Experiment," The Onion, April 17, 2017, https://www.theonion.com/cern-researchers-apologize-for-destruction-of-5-paralle-1819579830.

[10] Matthijs Maas and José Jaime Villalobos, "International AI Institutions: A literature review of models, examples, and proposals," (Legal Priorities Project, September 2023), https://dx.doi.org/10.2139/ssrn.4579773.

[11] Sam Clarke, apc, Jonas Schuett, "Survey on AI existential risk scenarios," AI Alignment Forum blog, June 8 2021, https://www.alignmentforum.org/posts/WiXePTj7KeEycbiwK/survey-on-ai-existential-risk-scenarios.

[12] Mustafa Suleyman and Eric Schmidt, "We need an AI equivalent of the IPCC," Finacial Times, October 10, 2023, https://www.ft.com/content/d84e91d0-ac74-4946-a21f-5f82eb4f1d2d.

[13] Vincent Manancourt and Tom Bristow, "British PM Rishi Sunak secures 'landmark' deal on AI testing," *Politico*, November 2, 2023, https://www.politico.eu/article/british-pm-rishi-sunak-secures-landmark-deal-on-ai-testing.

[14] Matthijs Maas and José Jaime Villalobos, "International AI Institutions: A literature review of models, examples, and proposals."

[15] James Broughel, "Creating an IPCC for AI Would Be a Historic Mistake," Forbes, November 10, 2023, https://www.forbes.com/sites/jamesbroughel/2023/11/10/creating-an-ipcc-for-ai-would-be-a-historic-mistake/?sh=415e50876d37.

[16] Lewis Ho et al., "International Institutions for Advanced AI," pre-print, July 11, 2023, https://arxiv.org/abs/2307.04699.