- **TECH**
- **ARTIFICIAL INTELLIGENCE**

Exclusive: U.S. Must Move 'Decisively' to Avert 'Extinction-Level' Threat From AI, Government-Commissioned Report Says

BY **BILLY PERRIGO**

MARCH 11, 2024 9:00 AM EDT

The U.S. government must move "quickly and decisively" to avert substantial national security risks stemming from artificial intelligence (AI) which could, in the worst case, cause an "extinction-level threat to the human species," says a report commissioned by the U.S. government published on Monday.

"Current frontier AI development poses urgent and growing risks to national security," the report, which TIME obtained ahead of its publication, says. "The rise of advanced AI and AGI [artificial general intelligence] has the potential to destabilize global security in ways reminiscent of the introduction of nuclear weapons." AGI is a hypothetical technology that could perform most tasks at or above the level of a human. Such systems do not currently exist, but the leading AI labs are working toward them and many expect AGI to arrive within the next five years or less.

The three authors of the report worked on it for more than a year, speaking with more than 200 government employees, experts, and workers at frontier AI companies—like OpenAI, Google DeepMind, Anthropic and Meta— as part of their research. Accounts from some of those conversations paint a disturbing picture, suggesting that many AI safety workers inside cutting-edge labs are concerned about perverse incentives driving decisionmaking by the executives who control their companies.

The finished document, titled "An Action Plan to Increase the Safety and Security of Advanced AI," recommends a set of sweeping and unprecedented policy actions that, if enacted, would radically disrupt the AI industry. Congress should make it illegal, the report recommends, to train AI models using more than a certain level of computing power. The threshold, the report recommends, should be set by a new federal AI agency, although the report suggests, as an example, that the agency could set it just above the levels of computing power used to train current cutting-edge models like OpenAI's GPT-4 and Google's Gemini. The new AI agency should require AI companies on the "frontier" of the industry to obtain government permission to train and deploy new models above a certain lower threshold, the report adds. Authorities should also "urgently" consider outlawing the publication of the "weights," or inner workings, of powerful AI models, for example under open-source licenses, with violations possibly punishable by jail time, the report says. And the government should further tighten controls on the manufacture and export of AI chips, and channel federal funding toward "alignment" research that seeks to make advanced AI safer, it recommends.

The report was commissioned by the State Department in November 2022 as part of a federal contract worth $250,000, according to public records. It was written by Gladstone AI, a four-person company that runs technical briefings on AI for government employees. (Parts of the action plan recommend that the government invests heavily in educating officials on the technical underpinnings of AI systems so they can better understand their risks.) The report was delivered as a 247-page document to the State Department on Feb. 26. The State Department did not respond to several requests for comment on the report. The recommendations "do not reflect the views of the United States Department of State or the United States Government," the first page of the report says.

The report's recommendations, many of them previously unthinkable, follow a dizzying series of major developments in AI that have caused many observers to recalibrate their stance on the technology. The chatbot ChatGPT, released in November 2022, was the first time this pace of change became visible to society at large, leading many people to question whether future AIs might pose existential risks to humanity. New tools, with more capabilities, have continued to be released at a rapid clip since. As governments around the world discuss how best to regulate AI, the world's biggest tech companies have fast been building out the infrastructure to train the next generation of more powerful systems—in some cases planning to use 10 or 100 times more computing power. Meanwhile, more than 80% of the American public believe AI could accidentally cause a catastrophic event, and 77% of voters believe the government should be doing more to regulate AI, according to recent polling by the AI Policy Institute.

Outlawing the training of advanced AI systems above a certain threshold, the report states, may "moderate race dynamics between all AI developers" and contribute to a reduction in the speed of the chip industry manufacturing faster hardware. Over time, a federal AI agency could raise the threshold and allow the training of more advanced AI systems once evidence of the safety of cutting-edge models is sufficiently proven, the report proposes. Equally, it says, the government could lower the safety threshold if dangerous capabilities are discovered in existing models.

The proposal is likely to face political difficulties. "I think that this recommendation is extremely unlikely to be adopted by the United States government" says Greg Allen, director of the Wadhwani Center for AI and Advanced Technologies at the Center for Strategic and International Studies (CSIS), in response to a summary TIME provided of the report's recommendation to outlaw AI training runs above a certain threshold. Current U.S. government AI policy, he notes, is to set compute thresholds above which additional transparency monitoring and regulatory requirements apply, but not to set limits above which training runs would be illegal. “Absent some kind of exogenous shock, I think they are quite unlikely to change that approach,” Allen says.

Jeremie and Edouard Harris, the CEO and CTO of Gladstone respectively, have been briefing the U.S. government on the risks of AI since 2021. The duo, who are brothers, say that government officials who attended many of their earliest briefings agreed that the risks of AI were significant, but told them the responsibility for dealing with them fell to different teams or departments. In late 2021, the Harrises say Gladstone finally found an arm of the government with the responsibility to address AI risks: the State Department’s Bureau of International Security and Nonproliferation. Teams within the Bureau have an inter-agency mandate to address risks from emerging technologies including chemical and biological weapons, and radiological and nuclear risks. Following briefings by Jeremie and Gladstone's then-CEO Mark Beall, in October 2022 the Bureau put out a tender for report that could inform a decision whether to add AI to the list of other risks it monitors. (The State Department did not respond to a request for comment on the outcome of that decision.) The Gladstone team won that contract, and the report released Monday is the outcome.

The report focuses on two separate categories of risk. Describing the first category, which it calls “weaponization risk,” the report states: “such systems could potentially be used to design and even execute catastrophic biological, chemical, or cyber attacks, or enable unprecedented weaponized applications in swarm robotics.” The second category is what the report calls the “loss of control” risk, or the possibility that advanced AI systems may outmaneuver their creators. There is, the report says, “reason to believe that they may be uncontrollable if they are developed using current techniques, and could behave adversarially to human beings by default.”

Both categories of risk, the report says, are exacerbated by “race dynamics” in the AI industry. The likelihood that the first company to achieve AGI will reap the majority of economic rewards, the report says, incentivizes companies to prioritize speed over safety. “Frontier AI labs face an intense and immediate incentive to scale their AI systems as fast as they can,” the report says. “They do not face an immediate incentive to invest in safety or security measures that do not deliver direct economic benefits, even though some do out of genuine concern.”

The Gladstone report identifies hardware—specifically the high-end computer chips currently used to train AI systems—as a significant bottleneck to increases in AI capabilities. Regulating the proliferation of this hardware, the report argues, may be the “most important requirement to safeguard long-term global safety and security from AI.” It says the government should explore tying chip export licenses to the presence of on-chip technologies allowing monitoring of whether chips are being used in large AI training runs, as a way of enforcing proposed rules against training AI systems larger than GPT-4. However the report also notes that any interventions will need to account for the possibility that overregulation could bolster foreign chip industries, eroding the U.S.’s ability to influence the supply chain.

The report also raises the possibility that, ultimately, the physical bounds of the universe may not be on the side of those attempting to prevent proliferation of advanced AI through chips. “As AI algorithms continue to improve, more AI capabilities become available for less total compute. Depending on how far this trend progresses, it could ultimately become impractical to mitigate advanced AI proliferation through compute concentrations at all.” To account for this possibility, the report says a new federal AI agency could explore blocking the publication of research that improves algorithmic efficiency, though it concedes this may harm the U.S. AI industry and ultimately be unfeasible.

The Harrises recognize in conversation that their recommendations will strike many in the AI industry as overly zealous. The recommendation to outlaw the open-sourcing of advanced AI model weights, they expect, will not be popular. “Open source is generally a wonderful phenomenon and overall massively positive for the world,” says Edouard, the chief technology officer of Gladstone. “It’s an extremely challenging recommendation to make, and we spent a lot of time looking for ways around suggesting measures like this.” Allen, the AI policy expert at CSIS, says he is sympathetic to the idea that open-source AI makes it more difficult for policymakers to get a handle on the risks. But he says any proposal to outlaw the open-sourcing of models above a certain size would need to contend with the fact that U.S. law has a limited reach. “Would that just mean that the open source community would move to Europe?” he says. “Given that it's a big world, you sort of have to take that into account.”

Despite the challenges, the report’s authors say they were swayed by how easy and cheap it currently is for users to remove safety guardrails on an AI model if they have access to its weights. “If you proliferate an open source model, even if it looks safe, it could still be dangerous down the road,” Edouard says, adding that the decision to open-source a model is irreversible. “At that point, good luck, all you can do is just take the damage.”

The third co-author of the report, former Defense Department official Beall, has since left Gladstone in order to start a super PAC aimed at advocating for AI policy. The PAC, called Americans for AI Safety, officially launched on Monday. It aims to make AI safety and security "a key issue in the 2024 elections, with a goal of passing AI safety legislation by the end of 2024," the group said in a statement to TIME. The PAC did not disclose its funding commitments, but said it has "set a goal of raising millions of dollars to accomplish its mission."

Before co-founding Gladstone with Beall, the Harris brothers ran an AI company that went through YCombinator, the famed Silicon Valley incubator, at the time when OpenAI CEO Sam Altman was at the helm. The pair brandish these credentials as evidence they have the industry’s interests at heart, even as their recommendations, if implemented, would upend it. “Move fast and break things, we love that philosophy, we grew up with that philosophy,” Jeremie tells TIME. But the credo, he says, ceases to apply when the potential downside of your actions is so massive. “Our default trajectory right now,” he says, “seems very much on course to create systems that are powerful enough that they either can be weaponized catastrophically, or fail to be controlled.” He adds: “One of the worst-case scenarios is you get a catastrophic event that completely shuts down AI research for everybody, and we don't get to reap the incredible benefits of this technology.”

Retrieved July 1, 2024 from AI Poses Extinction-Level Risk, State-Funded Report Says | TIME