StakeOut.AI

NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

Comment Submitted to the National Institute of Standards & Technology
February 1, 2024

StakeOut.AI, a nonprofit seeking to prevent economic disempowerment caused by artificial intelligence (AI), encourages the National Institute of Standards & Technology (NIST) to take a very strong regulatory approach with respect to AI.

AI replacement dynamics in the AI industry demonstrate that the industry should not be trusted to regulate itself.

One of the biggest challenges to mapping, measuring, and managing the trustworthiness of generative AI systems is that some leaders in the AI industry are at best indifferent to, and at worst explicitly desire, humanity's replacement by AI. This AI replacement includes not only the economic replacement of most human jobs — which would be an unprecedented challenge for U.S. democracy and its institutions — but also the replacement of humanity as a species. Behind AI industry leaders' pursuit of this AI replacement dynamic is a combination of two motives: the pro-AI ideological motive and the profit motive. This suggests that the U.S. government's status-quo approach of largely relying on the AI industry's voluntary commitments may be inadequate for protecting the national interest. In this comment, we at StakeOut.AI contend that the inadequacies of self-regulation can be mitigated by subjecting the AI industry to meaningful democratic oversight — such as President Biden's Executive Order on AI safety — as well as to U.S. leadership in a robust system of intergovernmental coordination on such oversight.

To illustrate the threat of AI industry figures' loyalties to AI replacement, consider University of Alberta professor Rich Sutton: an AI research pioneer with deep ties to the U.S. AI industry. Professor Sutton is well-known for being the first-ever advisor of DeepMind, now Google DeepMind. He is also well-known for pioneering the AI method of reinforcement learning: a key method where AI systems teach themselves instead of being taught by humans. On July 7, 2023, Professor Sutton gave an academic talk at the World Artificial Intelligence Conference in Shanghai, entitled "AI Succession," denoting the replacement of the human species by an AI 'successor species.'[1] In it, Professor Sutton outlined his belief that humans will be inevitably replaced by AI, and that we humans should not resist, but embrace AI succession and prepare for it. He then outlined a concrete research plan to create an AI successor species (which he called the Alberta Plan for AI research) and asked for funding. Instead of being

shunned for his controversial goal of replacing the human species, Professor Sutton found a funding partner to create superintelligent AI. He was onboarded to Keen Technologies: a startup founded by decorated programmer John Carmack to create superintelligent AI.[2]

Professor Sutton is not the only AI industry figure who is ideologically predisposed to AI replacement. Another example is Google co-founder Larry Page. Back when Larry Page was still the CEO of Google, he had a tense conversation with Elon Musk at Musk's Napa Valley birthday party. TIME reports that when Musk warned that AI systems might replace humans unless we build safeguards, Page replied something to the tune of "Why would it matter if machines someday surpassed humans in intelligence, even consciousness? It would simply be the next stage of evolution."[3] Since then, Google DeepMind has scaled up to become one of the world's leading AI companies.

Google DeepMind is arguably behind only OpenAI (the creator of ChatGPT) in the race to develop superintelligent AI. OpenAI has the mission of creating "highly autonomous systems that outperform humans at most economically valuable work."[4] Such superintelligent AI systems are called Artificial General Intelligence — or AGI for short — because the practical capability advantage these hypothetical systems have over humans would be general, rather than limited to a specific domain. Meta — another player in the AGI race — goes one irresponsible step further than OpenAI, in that Meta plans to create and release AGI to be downloadable by essentially anyone with an Internet connection.[5] At the current level of progress in AI safety research, even state-of-the-art guardrails on widely downloadable superintelligent AI models are non-robust; they can be easily removed, so that the user can modify the model for their own purposes.[6] And criminals, rogue actors, and foreign adversaries would be among those who would be able to download the model weights of such a superintelligent AI system released by Meta.

The silver lining is that the companies currently leading the race to create superintelligent AI are all based in the U.S. (with the partial exception of Google DeepMind, the core of which was formed when the U.K. AI startup DeepMind was bought by the U.S. company Google). Consequently, the U.S. government has many promising policy levers that can help shape this technological race in the direction of preventing catastrophic failures and ensuring societal benefits via increased system safety.

Controlling superintelligent AI is an unsolved scientific problem.

AI companies are racing to be first-to-market with superintelligent AI systems. But among the world's leading AI scientists who do not work in the AI industry (and thus tend to

have less of a financial conflict of interest), many agree that the scientific problem of how to ensure that superintelligent AI remains under human control remains presently unsolved.

Consider Professor Geoffrey Hinton, the most-cited AI researcher of all time. He won the Turing Award for pioneering the method of deep learning, arguably the most important method in developing modern AI systems like ChatGPT. While Professor Hinton used to work on such AI systems at Google, he left the company in order to warn about the dangers posed by the very AI technology he pioneered.[7] In the short term, Professor Hinton's concern was that AI-generated images, text, and deepfake videos would prevent people from being able to distinguish truth from falsehoods, misinformation, and impersonations; and that AI would replace many human jobs. These predictions are holding up well, with current generative AI systems inflicting substantial harms, such as the manipulation of elections,[8] fraud,[9] and nonconsensual deepfake pornography[10]; as well as the scraping of copyrighted work to train AI models, using the resulting content to compete against the very same people who created the work the models scraped.[11,12] These harms will be increasingly difficult to mitigate, given Professor Hinton's prediction that AI systems will become smarter than humans. In an interview with CNN journalist Jake Tapper, Professor Hinton warned, "If it gets to be much smarter than us, it will be very good at manipulation because it would have learned that from us. And there are very few examples of a more intelligent thing being controlled by a less intelligent thing."[13]

Since then, AI scientists have found that even current AI systems can learn from their training the ability and tendency to manipulate humans, even if the creators of these AI systems had not intended for these systems to be manipulative and deceptive.[14] It is important to note that at the present, people being systematically manipulated by the decisions of an AI system (as opposed to deepfake videos, audio, or text produced by a human prompter using AI) is not yet a major threat vector. But this may change — potentially in the near future — due to AI companies' race to be first-to-market to superintelligent AI.

The scientific problem of how to control superintelligent AI remains unsolved.[15] AI has a black box problem, where its internal decision-making processes are opaque, even to its human creators, monitors, and auditors. This issue is compounded by the rapid advance of AI capabilities, which is substantially outpacing our understanding of their internal decision-making processes (via an internals-based safety test). This makes it so that if an AI system unexpectedly learns deceptive tendencies during its training, its human creators likely will not be able to successfully probe for these deceptive tendencies in advance. Indeed, of the two potential solutions to probe for deceptive tendencies — internals-based safety tests and behavior-based safety tests — neither are robust. Regarding the former, it is currently unknown how to robustly find or rule out deceptive tendencies (or any other complex trait) by inspecting the AI system's

internal decision-making process, due to the black box problem. Regarding the latter, a sufficiently advanced AI system would be able to cheat the behavior-based safety test, analogous to how a student can cheat on an exam and thereby pass it even though they have not learned the material that the exam is designed to test.[14]

This raises the concern that AI safety technicians may be unequipped to prevent a sufficiently advanced AI system from deceiving humans, either following its deployment or its escape from the training environment. The AI system would then be free to pursue its own goal, even if this goal is unintended by its human creators. Such a development would create or magnify several high-risk AI threat vectors, such as goal-directed social influence campaigns, goal-directed cyberattacks, goal-directed engineering of pandemics, goal-directed use of Lethal Autonomous Weapons (LAW), autonomous replication, and autonomous power-seeking.[16]

Mitigating AI risks via robust standards

We at StakeOut.AI propose that the AI industry should be subject to robust standards of risk management, just like other industries that can expose the public to high risk if left unregulated. The standard way of practicing risk management — as practiced in industries like nuclear power, aviation, and pharmaceuticals — requires each company to quantitatively demonstrate that the risk in question is below the pre-agreed threshold.

When predicting the degree to which quantitative risk bounds can robustly ensure safety, it is important to consider the degree to which the system in question is understood at the current level of scientific knowledge. In domains where experts understand a high proportion of the knowledge relevant to the given system — such as situation-specific behavior, internal mechanisms, and catastrophic failure modes — quantitative risk bounds lead to a high likelihood that the system's risks will be managed. However, in domains like generative AI where experts understand a low or even near-zero proportion of the knowledge relevant to the given system, quantitative risk bounds will need to be applied in addition to a principle of not ruling out unknown risks: in order for the system's risks to have a decent chance at being managed.

Thus, it is important for the AI industry's risk-management practitioners and standard-setters to not be overconfident in the current scientific understanding of system-relevant knowledge. It is also important to prioritize increasing the degree to which system-relevant knowledge is scientifically understood going into the future, such as by societally incentivizing the necessary scientific research via various measures.

In addition to quantitative risk bounds, there are other highly important aspects of robust risk management. StakeOut.AI — as part of the Future of Life Institute's AI governance proposal presented by Max Tegmark at the UK AI Safety Summit — researched a 'scorecard' comparing several AI governance proposals. Our researched scorecard[17] scored these AI governance proposals based on whether they satisfied certain important criteria: such as 'quantitative risk bounds.' A given proposal was scored to have satisfied this criterion if it "defines numerical thresholds or limits pertaining to the potential risks or harm an AI system might pose. In addition to this, we also scored the proposals with respect to the following criteria:

'Burden of proof on developer to demonstrate safety?'
This criterion is satisfied if the proposal "obliges AI developers to proactively provide evidence or justification of the safety of their systems prior to training and deployment."

'Compute limits?'
This criterion is satisfied if the proposal "sets boundaries on the computational resources or power that an AI system can use."

'Liability requirements?'
This criterion is satisfied if the proposal "outlines the responsibilities and legal consequences for developers or users should their AI system cause harm or operate outside of its defined parameters."

'Third-party safety audit requirements?'
This criterion is satisfied if the proposal "stipulates that AI systems must undergo a systematic and independent examination to ensure safety measures are met."

'Registration requirements?'
This criterion is satisfied if the proposal "mandates the recording and submission of specific details about an AI system prior to its training and deployment."

'Doesn't exempt open source?'
This criterion is satisfied if the proposal "does not exempt open-source or widely released AI models from the requirements of the proposal."

'Doesn't exempt LLMs?'
This criterion is satisfied if the proposal "does not exempt large language models from the requirements of the proposal."

‘Doesn’t exempt military AI?’
This criterion is satisfied if the proposal “does not exempt military training and deployment of AI systems from the requirements of the proposal.”

We at StakeOut.AI propose that the U.S. government can achieve significant societal benefits by improving risk-management practices with respect to these listed criteria: such as by ensuring that the companion resource to the AI Risk Management Framework adheres to each of these criteria. We also propose that by doing so, the U.S. government can advance its leadership role in the intergovernmental pursuit of getting the AI industry to adopt standard risk-management practices, which will likely provide significant societal benefits by meaningfully increasing industry safety and mitigating the risks of catastrophic failures.

Finally, it is important to note that the U.S. has already committed to a human-centered AI framework which states that AI should be transparent and drive inclusive growth. Specifically, the U.S. has endorsed the Organization for Economic Cooperation and Development (OECD) Principles on Artificial Intelligence. Principle 1.2 reads: “AI actors should respect the rule of law, human rights and democratic values throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights.”[18] Dignity, fairness, and social justice require presciently regulating the AI industry, so as to prevent the risks of disempowerment and catastrophic tail risks that the AI industry would otherwise cause.

Endnotes

[1] Rich Sutton, *AI Succession*, World Artificial Intelligence Conference (July 7, 2023), available at https://www.youtube.com/watch?v=NgHFMolXs3U.
[2] Lynda Vang, Richard S. Sutton, and Cam Linke, *John Carmack and Rich Sutton Partner to Accelerate Development of Artificial General Intelligence*, ALBERTA MACHINE INTELLIGENCE INSTITUTE (Sept. 25, 2023), https://www.amii.ca/latest-from-amii/john-carmack-and-rich-sutton-agi/.
[3] Walter Isaacson, *Inside Elon Musk's Struggle for the Future of AI*, TIME (Sept. 6, 2023), https://time.com/6310076/elon-musk-ai-walter-isaacson-biography/.
[4] *OpenAI Charter*, OPENAI (April 9, 2018), https://openai.com/charter.
[5] John Koetsier, *Meta To Build Open-Source Artificial General Intelligence For All, Zuckerberg Says*, FORBES (Jan. 18, 2024), https://www.forbes.com/sites/johnkoetsier/2024/01/18/zuckerberg-on-ai-meta-building-agi-for-everyone-and-open-sourcing-it/.
[6] Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin, *Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models*, ARXIV (Oct. 4, 2023), https://arxiv.org/abs/2310.02949.
[7] Cade Metz, *'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead*, NEW YORK TIMES (May 4, 2023), https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html.
[8] Sander van der Linden, *AI-Generated Fake News Is Coming to an Election Near You*, CNN (Jan 22, 2024), https://www.wired.com/story/ai-generated-fake-news-is-coming-to-an-election-near-you/.
[9] Nabila Ahmed, Adam Haigh, Ainsley Thomson, and Ellie Harmsworth, *Deepfake Imposter Scams Are Driving a New Wave of Fraud*, BLOOMBERG (Aug 23, 2023), https://www.bloomberg.com/news/articles/2023-08-21/money-scams-deepfakes-ai-will-drive-10-trillion-in-financial-fraud-and-crime.
[10] Jeremy Kahn, *Taylor Swift Deepfake Porn Points to a Fundamental Problem: AI Can Make It, But Can't Police It*, FORTUNE (Jan. 30, 2024), https://fortune.com/2024/01/30/ai-policing-taylor-swift-porn-deepfakes-porn/.
[11] Winston Cho, *Top Authors Join Lawsuit Against OpenAI Over "Mass-Scale Copyright Infringement" of Novels*, THE HOLLYWOOD REPORTER (Sep. 20, 2023), https://www.hollywoodreporter.com/business/business-news/top-authors-join-lawsuit-against-openai-over-mass-scale-copyright-infringement-of-novels-1235595123/.
[12] Michael M. Grynbaum and Ryan Mac, *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*, NEW YORK TIMES (Dec. 27, 2023), https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.
[13] Jennifer Korn, *Why the 'Godfather of AI' decided he had to 'blow the whistle' on the technology*, CNN (May 3, 2023), https://www.cnn.com/2023/05/02/tech/hinton-tapper-wozniak-ai-fears/index.html.

[14] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks, *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, ARXIV (Aug 28, 2023), https://arxiv.org/abs/2308.14752.
[15] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt, *Unsolved Problems in ML Safety*, ARXIV (Jun 16, 2022), https://arxiv.org/abs/2109.13916.
[16] Dan Hendrycks, Mantas Mazeika, Thomas Woodside, *An Overview of Catastrophic AI Risks*, ARXIV (Oct. 9, 2023), https://arxiv.org/abs/2306.12001.
[17] AI Governance Scorecard and Safety Standards Policy: Evaluating proposals for AI governance and providing a regulatory framework for robust safety standards, measures and oversight, FUTURE OF LIFE INSTITUTE (Oct. 30, 2023), *https://futureoflife.org/document/fli-governance-scorecard-and-safety-standards-policy/.*
[18] *Recommendation of the Council on Artificial Intelligence*, OECD (May 21, 2019), OECD/LEGAL/0449.