Machine Intelligence Research Institute
Berkeley, California

Point of contact:
Lisa Thiergart and Peter Barnett
Technical Governance Team
techgov@intelligence.org

May 17th, 2024

The Office of Senator Mitt Romney
United States Senate
Washington, DC

Re: Framework to Mitigate AI-Enabled Extreme Risks

Our organization, the Machine Intelligence Research Institute (MIRI) is a research nonprofit based in Berkeley, California, founded in 2000. We focus on research and analysis regarding the challenges humanity will face in safely navigating the transition to a world with smarter-than-human artificial intelligence (AI). If humanity can safely navigate the emergence of these systems, we believe this will lead to unprecedented levels of prosperity. However, as our CEO, Malo Bourgon, outlined in his written statement as part of his participation in a Senate AI Insight Forum, if not handled with extreme care, we believe that there is a significant chance that the development of such systems will pose an existential risk to humanity.

We were heartened to see the announcement of the framework by Senators Romney, Reed, Moran, and King, which we believe is one of the first legislative frameworks put forward in Congress that is directly grappling with how to mitigate some of the extreme risks posed by advanced AI systems that may be just over the horizon. We were also pleased to see the emphasis on the importance of international coordination, in the letter to Leader Schumer, and Senators Rounds, Heinrich, and Young, which we view as essential for mitigating the extreme risks posed by these systems.

This framework is an excellent first step, but we believe that it could be strengthened by broadening its scope to not only attempt to address the extreme risks posed by misuse of such systems, but also risks arising from the challenges of controlling the behavior of the systems themselves as they become increasingly capable. That is, the risk that as these general purpose AI systems begin to match and vastly surpass human performance at most cognitive tasks, their use by even well-meaning actors could result in catastrophic consequences. Below we'll expand a little on the source of these risks (though a proper explanation is out of scope for this response), as well as suggest some ways that this framework could be modified as a first step to start addressing such risks.

Overview of risk from model autonomy

There are many examples of contemporary AI systems behaving in harmful ways their designers didn't intend. For example, in 2023, Microsoft's Bing Chat produced offensive and inappropriate content, including issuing threats, in conversations with users.[1] Microsoft clearly did not intend or anticipate that the model would behave in this way, and while the harms were minor, this serves as an illustrative example of how challenging it can be to predict how these types of systems will behave.

These challenges arise due to the general purpose nature of these systems in combination with how little we understand about their inner workings (including which internal goals and motivations they may have developed), which makes it very difficult (if not impossible with current levels of understanding) to predict how they may behave in novel situations and environments. These types of failures have so far still been limited in scope due to limits of current AI models' capabilities. As models grow more capable and able to act with increased autonomy,[2] the harms caused by such misbehavior are likely to become much more consequential.

These are risks acknowledged by current frontier AI developers, e.g., they are included in the scaling policies released by Open AI and Anthropic in which they refer to them as *risks from model autonomy*.[3] [4] Both labs have made voluntary commitments to monitor and evaluate for these risks, concrete examples of which include, a model launching a cyberattack to obtain further resources or information, attempts at *self-exfiltration*[5], and deliberately manipulating human operators in service of pursuing the model's own goals. Importantly, risks from model autonomy can occur despite a user being well-intentioned and well-informed. Further, to immediate damages from model autonomy, insufficient monitoring for risks from model autonomy can increase the likelihood of loss of control over an AI system.[6]

Since these risks arise from the behavior of the models themselves, they are not just relevant during deployment, but may also occur during training or internal use and testing. Therefore, it is important to conduct regular model evaluations focused on model autonomy from the training stage onward.

---

[1] Perrigo, Billy. "The new AI-powered Bing is threatening users. That's no laughing matter." *Time Magazine* (2023).
[2] Wang, Lei, et al. "A survey on large language model based autonomous agents." *Frontiers of Computer Science* 18.6 (2024): 1-26.
[3] OpenAI. "Preparedness Framework (Beta)" (2023)
[4] Anthropic. "Anthropic's Responsible Scaling Policy Version 1.0" (2023)
[5] OpenAI. "Preparedness Framework (Beta)" (2023)
[6] Bengio, Yoshua, et al. "Managing AI risks in an era of rapid progress." *arXiv preprint arXiv:2310.17688* (2023).

Evaluating for Risks from Model Autonomy

Similarly to evaluations for misuse risks, accurately evaluating for model autonomy risks can be extremely challenging. For example, current misuse evaluations rely on consulting human experts to identify the most dangerous actions and then testing to see if the model or the model together with a non-expert human can succeed in taking the dangerous action. This might look like asking a human to use the AI system to find cyber vulnerabilities or provide advice for producing biological weapons. Whereas misuse evaluations must determine the range of capabilities a human can elicit from the model, an autonomy evaluation faces the much harder challenge of determining the full range of capabilities of a model overall.

A challenge both these evaluations face, is that there is missing scientific understanding of how to measure success at eliciting the model's full range of capabilities. This concept is often referred to as *capability elicitation*. It is likely that significant novel research and development will be required before AI developers have the ability to demonstrate a sufficiently high degree of understanding of the full-range capabilities of their models and ability to control them.

This difficulty should not mean the necessity of measuring risks from model autonomy remain neglected, but rather should motivate further research and scientific inquiry in advance of further capability acceleration.

Concrete Recommendations

Our primary recommendation for the framework is to include requirements to evaluate AI models for risks from model autonomy.

Evaluations should be performed during development as well as before deployment. This is because model autonomy risk arises from the model itself, not from humans misusing it. As AI systems become more powerful, they may pose risks during their training or internal use and testing.

Cybersecurity standards should be tailored to prevent risks from model autonomy as well as risks from human attackers. Whereas a lot of focus in current cybersecurity industry practices is placed on preventing model weights being leaked or stolen, we believe additional focus should be placed on preventing risks from model autonomy. For example, these measures may be required to prevent self-exfiltration, where a model may transfer its own weights to a less secure location. Fortunately, at least initially cybersecurity measures intended to prevent models being leaked or stolen will likely help to also defend against risks from model autonomy. However, although current models are unlikely to possess this capability, we believe specialized cybersecurity infrastructure can be costly and time-intensive to implement, and that industry standards should

consider more stringent measures well in advance of autonomy risks being directly measured.

Additional Comments

Compute thresholds should account for algorithmic improvements. The current framework uses a threshold of 10^{26} operations, which is the threshold also used in Executive Order 14110. However, because of algorithmic progress each year it becomes possible to train models of the same capabilities with less compute.[7] The compute required to train a model of a given capability halves approximately every 8 months. Because of this we believe the compute threshold will likely need to be adjusted downward over time.

The specific numbers used for current compute thresholds are initial heuristics; we believe that it's essential for the relevant regulatory body to have the authority to nimbly update this number as more understanding is developed. Ideally, these compute thresholds will ultimately be replaced by thresholds based directly on model capabilities. But currently, it is not possible to know what capabilities a model will have until after it has been trained.

Safety fine-tuning is not sufficient. We would like to emphasize that fine-tuning powerful AI models not to display dangerous behavior is not sufficient to ensure safety. Such fine-tuning can quickly and cheaply be undone with additional fine-tuning. This means that if model weights are intended to be made publicly available, the model should be treated as if there is no safety fine-tuning. This is also a risk if the model weights are unintentionally leaked or stolen. Additionally, even without access to model weights, it may be possible to circumvent safety fine-tuning by jailbreaking the model. More effective mitigations would include monitoring of model inputs and outputs, while remaining mindful that such monitoring may be imperfect.

External safety and security audits are essential. AI labs should not be responsible for grading their own homework. As such, we recommend that the framework should explicitly require external audits, which may be performed by the oversight body or another appropriate organization. This should include evaluations for the safety of AI systems; testing the model's capabilities and whether it would be able to cause harm in the environments in which it operates. This should also include security audits, to ensure AI developers are adhering to the required cybersecurity standards.

[7] Ho, Anson, et al. "Algorithmic progress in language models." *arXiv preprint* *arXiv:2403.05812* (2024).

Closing

This framework is a promising initial step in legislating to prevent extreme risks from advanced AI systems. We believe the practical implementation of this framework will matter a lot in how successfully it mitigates the stated risks; both the risks from human misuse and from unintended behavior of powerful models. It seems particularly important to strike a correct balance between a strong implementation that effectively mitigates these risks while not imposing an unnecessary burden on developers or users. By default, incentives may push towards weak implementation that provides an illusion of safety without defending against the core risks.

We'd be very happy to serve as a resource for your offices regarding the technical and further conceptual implementation of this promising framework, or more generally to further discuss the challenges of successfully mitigating the risks of future AI systems. Please feel free to contact us at techgov@intelligence.org.