Machine Intelligence Research Institute
Berkeley, California

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

April 29th, 2024

Office of Management and Budget
Washington DC

Re: OMB Procurement of Artificial Intelligence RFI (OMB-2024-0004-0001)

Our organization, the Machine Intelligence Research Institute (MIRI), is focused on increasing the probability that humanity can safely navigate the transition to a world with smarter-than-human AI. We investigate ways to mitigate the catastrophic and societal-scale risks associated with artificial intelligence (AI) systems as they near and surpass the capabilities of humans. In the context of this memo, we will primarily focus on "safety-impacting" AI systems, as these systems are most capable of causing societal-scale damage.

AI systems are rapidly becoming more powerful and are able to automate increasingly more physical and intellectual labor. Government agencies will want to use and procure these powerful systems. We urge the OMB to be forward-looking, such that requirements and regulations put in place today are also adequate to address the challenges of future AI systems.

When we discuss future powerful AI systems, we are considering systems like the following examples:

- Systems which are capable of automating science and R&D, such as work which is performed at DoE or DoD, or other government research laboratories.
- AI which is in control of critical infrastructure and is capable of making complex decisions.
- AI, which if stolen by a malicious actor, could be repurposed to create chemical, biological, radiological, nuclear (CBRN) threats.

We want to stress that we are focused on powerful "frontier" AI systems, and that the risk management framework and suggestions we discuss need not be applied to systems which do not pose large-scale harms. Many AI systems will be safe, either because they only have narrow capabilities (for example, a weather prediction AI cannot do any other tasks) or because they are incapable (for example, non-frontier language model AI systems).

There are not currently good methods for establishing whether an AI system is dangerous. Currently we recommend treating AI systems trained with more than 10^{26} floating point operations as potentially dangerous, as is in line with the reporting requirements from Executive Order 14110.[1] We would also encourage research into better measurement and classification of AI capabilities.

In our response to this Request for Information, we want to focus on having good risk assessment and management processes as part of the procurement process, while also being careful to be clear about which risks can and cannot be measured. In general, we urge a cautious approach, focused on having positive evidence that AI systems are safe in order to avoid bad outcomes.

The capabilities of large AI models today are unpredictable, it can be extremely challenging to determine if an AI model is capable of a task, both before training and after the system has been trained. Historically, there have been many times when new methods were developed long after the initial training which revealed an AI system to be more capable than originally thought. The current method for testing an AI system's capabilities is using "behavioral evaluations", which involve a human prompting an AI in an attempt to get it to perform specific proxy tasks. It is hoped that if the AI does not succeed at the proxy task, it will be unable to perform the dangerous task we care about. For example, if an AI is unable to perform well at a certain programming test, it may be assumed to be incapable of assisting in cyber-offence attacks. These evaluations form a large part of the voluntary commitments made by large AI developers OpenAI and Anthropic, although these have also been criticized for various limitations.

It appears likely that risk assessment and management of powerful AI systems will likely make use of these behavioral evaluations. For these to allow confidence in the safety of AI systems, certain difficulties must be overcome:

- It is difficult to elicit the maximum capabilities of an AI model, and to know whether capabilities have or have not been elicited.
- It is difficult to know if these proxy tasks truly measure the skills we care about.
- It is difficult to know if we have appropriately enumerated ways in which the AI system may cause harm, given how it will be deployed.

Overall, when procuring AI systems, the government should thoroughly evaluate the systems for ways in which they could cause unexpected harm. The agencies using powerful AI systems should be sure to have adequate cyber and organizational security in order to prevent these systems from being stolen by malicious actors, who could potentially use them for developing biological weapons or cyber attacks. This security should also be sufficient to prevent risks from "self-exfiltration", where an autonomous AI

[1] We believe this threshold should be lowered with time, as "algorithmic efficiency" improvements allow AI systems with the same capabilities to be trained with fewer computational resources.

system secretly transfers its weights and code outside of the secure facility. Again, we emphasize that only a few AI systems pose these risks; but those that do should be thoroughly evaluated for their ability to cause harm. These risks should be mitigated or the AI system should not be deployed.

We propose that during procurement, a risk assessment and management approach process should be followed which is able to adequately answer the following questions:

- In which ways could the AI system cause societal-scale harm?
    - Why are we confident that we have covered the space of ways in which this AI system could cause harm?
- Why are the evaluations and proxy tasks adequate for measuring the AI system's dangerous capabilities?
- How are we confident that we are actually measuring close to the AI's maximum capabilities for the specific tasks?

If these questions cannot be adequately answered, then we believe that it is not responsible for the government to use the AI system. We believe that similar restrictions should also be applied to private AI companies, in order to prevent large-scale harm from AI systems.

We believe that these requirements (and all of the requirements in the memo) should apply to all powerful AI used within government, including "National Security Systems" and AI used by the Intelligence Community.

Below we respond to specific questions in the RFI.

Comments in Response to Question 5: Materials provided by AI vendors to demonstrate compliance

Our answer for this question will pertain to only powerful AI systems, initially defined as AI systems trained using more than 10^{26} floating point operations.

We believe that an adequate risk assessment should be performed before a powerful AI system is procured, and that vendors should provide all appropriate materials to enable this. Initial ideas for risk assessments of powerful AI systems may be found here.

This risk assessment may be based on mapping ways in which the AI system could cause harm, and attempting to determine the AI system's capabilities. This includes the AI system causing harm due to being used by a malicious actor, as well as societal-scale harm caused by the AI system acting autonomously.

The vendor should assist in this assessment, and provide documentation and evidence to adequately answer the following questions:

- In what ways could the AI system cause harm, given the deployment context? This includes both harm from misuse by malicious actors and harm from the AI system acting autonomously.
    - Why can we be confident that this comprehensively maps the ways in which the AI system could cause harm?
- What capabilities can we measure to test the AI system's ability to cause harm?
    - What is the justification that these tests will measure the dangerous capabilities we care about?
- How are these capabilities elicited?
    - What is the justification that the AI system's capabilities were maximally elicited?

Answering these questions must not be a box-checking exercise, and if these questions are unable to be answered adequately, and in a principled manner, then it is not responsible to use or procure the AI system. Without adequate answers to these questions, a government agency cannot be confident that the use of the AI system will not cause harm.

Given the current state of AI evaluations, it may not be possible to answer these questions, and so for sufficiently powerful AI systems, it may not be responsible to use them.

In Section 5(d)(vii) ("Responsibly Procuring Generative AI"), the memo states that agencies should include risk management requirements in contracts for generative AI. **We believe that this should be mandatory for sufficiently powerful AI models (as defined above).** This section also mentions "adequate testing and safeguards", it is not possible to design safeguards without knowing an AI system's capabilities. **Hence, we believe that providing principled answers to the questions above should be a mandatory part of the risk management requirements for powerful AI systems.**

Comments in Response to Question 6: Testing and evaluation of AI systems by AI vendors and agencies

For powerful AI systems, we believe that risk and impact assessments should be performed independently by both the vendor and the agencies. The vendor will likely have more technical knowledge and experience, while the agencies will have stronger incentives to thoroughly map out the potential risks from powerful AI systems. Without independent risk assessment, vendors may face incentives to downplay the potential risks.

A vendor's assessment should be at least as thorough as the assessment performed by the agency. If the vendor's assessment is found to be significantly more lax, then this should be a strong reason not to procure powerful AI from the vendor.

The current techniques for risk assessment and evaluation for powerful AI are still new, and likely not adequate for future powerful AI systems (such as AI systems capable of automating large amounts of R&D). Vendors (and government agencies) should attempt to improve the state of risk assessment and evaluations of AI systems, to be able to fully predict the capabilities of AI systems and understand how they work in a principled way. This may be the only way to provide adequate assurance of safety for such systems.

Comments in Response to Questions 9 and 10, answered jointly: Reducing harm and promoting positive outcomes from AI systems

In our answer, we will focus on "safety-impacting" AI systems which are capable of causing societal-scale damage. Such systems may be capable of causing extreme harm to individuals (including killing people). Malicious actors could use AI for cyber attacks, or generate propaganda, as well as manipulative and deceptive content.

As stated previously, we believe we need robust risk assessment and management, and that this may not be possible given the current state of AI evaluations.

Below we discuss proposed changes to Section 5 ("Managing Risks from the Use of Artificial Intelligence").

Before procuring a powerful AI system, agencies should implement adequate risk management practices, or else they should not procure the AI.

5(c)(a)(iv)(A): Complete an AI impact assessment including the potential risks
We believe that this impact assessment should be shared with stakeholders, as long as this is safe (for example, as long as this does not involve sharing sensitive information). For systems which can control critical infrastructure or may pose societal-scale risks, the stakeholders may be the general public. In this case the impact assessment should be made publicly available.

5(c)(a)(iv)(B): Test the AI for performance in a real-world environment context
This real-world testing may require extensive safeguards (both to protect from malicious actors and to prevent unwanted actions from a powerful autonomous AI). These safeguards will likely be necessary in order to safely perform these tests. The safeguards used during testing may also be required after testing, and so the cost of continued safeguards should be factored into the cost of procuring the AI system.

This section states "Testing conditions should mirror as closely as possible the conditions in which the AI will be deployed". **We would suggest changing this wording to explicitly state that if the testing is inadequate or if safety cannot be**

guaranteed, then the AI system should not be used or procured. This may be the case for powerful AI which is capable of automating R&D, where it may currently be extremely difficult to ensure that the testing conditions match all of the real world scenarios the AI will encounter.

5(c)(a)(iv)(C): Independently evaluate the AI
We applaud this as an essential component for ensuring that powerful AI systems do not cause significant harm. However, as stated previously, there are inadequacies with current AI evaluations. If methods for evaluating and understanding AI systems do not improve, it may not be appropriate to use or procure powerful AI systems.

5(c)(a)(iv)(D, E, F): Conduct ongoing monitoring, Regularly evaluate risks from the use of AI, Mitigate emerging risks to rights and safety
Similarly, we applaud these as essential steps in ensuring AI is used safely. **We stress the need to have these systems in place before powerful AI systems are procured.**

There is a serious risk that malicious actors (including nation state actors) may attempt to steal powerful AI systems. It is essential that agencies using powerful AI in a manner in which it could be stolen (for example, if the agency was running a copy of the AI, rather than accessing it via an API), should have cybersecurity which is sufficient to defend against extremely capable threat actors.

5(c)(a)(iv)(G, H): Ensure adequate human training and assessment, Provide additional human oversight, intervention, and accountability as part of decisions or actions that could result in a significant impact on rights or safety
These again are important practices for ensuring safety. It may become extremely challenging or even infeasible for humans to test or assess the outputs from powerful AI systems. For example, if an AI is performing automated R&D in a domain the human is unfamiliar with, or if the AI is optimizing for its outputs to deceive the human.

We further note that there will likely be strong pressures to allow powerful AI systems to automate processes without adequate supervision; for example human supervision may be expensive or time-consuming. We recommend that risk management plans explicitly note this as a risk factor, and make explicit efforts to avoid this. We would recommend a standard where humans should understand all of the decisions made by AI systems; although we also note that it may be easy for humans to be fooled into thinking they understand when they don't.

5(c)(a)(iv)(I): Provide public notice and plain-language documentation
We strongly approve of this, and believe that this could work with 5(c)(a)(iv)(A) ("Complete an AI impact assessment"), where this impact assessment should be included in the plain-language documentation, and be available for appropriate oversight and public scrutiny.

We now comment on Section 5(d) ("Managing Risks in Federal Procurement of Artificial Intelligence").

5(d)(ii): Transparency and Performance Improvement

We are excited to see the steps listed in this section of the memo for their ability to help provide transparency and guide risk management. Here we will lay out ways in which these could be further improved.

- A (Obtaining adequate documentation to assess the AI's capabilities): There are serious limitations for assessing the capabilities of powerful AI systems given current methods. While access to the model, data, and system cards may be helpful here, these are currently not sufficient to truly assess the capabilities of powerful AI systems. For sufficiently powerful AI systems, without advances in methods to determine an AI system's capabilities, it may not be responsible to procure or use such a system.

- B (Obtaining documentation on the known limitations of the AI, and guidance on how the system should be used): We agree it is important for the known limitations of AI systems to be documented to the procuring agency. We note that there are often many additional limitations of systems which are not initially known and may be consequential, for example "jailbreaks" of current systems.

- D (Regularly evaluating claims made by contractors, including risk management measures): We thoroughly support this in spirit. We are however concerned that given the limitations of current AI evaluations, such evaluations may offer a false sense of security and safety. We suggest that for powerful AI systems, the risk management process should be required to answer the questions discussed in our answer to Question 5.

- E (Considering contracting provisions that incentivize the continuous improvement of procured AI): If there is continuous improvement of AI systems, this brings novel risks, such that safeguards may become inadequate. **If the AI system is being improved or modified, we believe that the risk management practices (including safeguards and evaluations) must also be continually updated.** We note that these improvements may come from the vendor or agency modifying the AI, or from the AI autonomously improving. Both these cases require up-to-date risk management, and this will likely be more difficult in the case of autonomous improvement as it may not be clear in which ways the AI system was modified.

5(d)(v): Overfitting to known test data

We emphasize that some of the failures of current AI evaluations can be seen as "overfitting to known test data". If an AI system is shown to fail at a specific task (for example, finding a certain vulnerability in a piece of software) it is important not to "overfit" to this test and assume that the AI is overall incapable of performing similar tasks (such as assisting in other cybersecurity attacks). We stress how important it is not to overgeneralize from the results of simple tests.

5(d)(vii): Responsibly procuring generative AI

Above, when we have discussed “powerful AI” (as defined using a floating point operation threshold), we believe that these systems are likely to be generative AI systems. Our recommendations and concerns as discussed above remain relevant in this context, and should be applied.

We will briefly comment on the specific risk management requirements listed in this section:

- A (“requiring adequate testing and safeguards”): We emphasize that adequate testing may currently be infeasible given current methods, and that it may be challenging to determine whether safeguards are adequate.
- B (“requiring results from internal or external testing and evaluation”): We again emphasize the limitations of current AI evaluations, and that it may not currently be possible to use these to gain confidence in the safety of a powerful AI system.
- D (“Incorporating relevant NIST standards”): We refer to "AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models” for an overview of applying relevant NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) to powerful AI systems. In particular, we would emphasize the following sections from the NIST RMF, although other sections are also important:
    - *Map 5.1: Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.*
    - *Measure 1.1: Approaches and metrics for measurement of AI risks enumerated during the “Map” function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not–or cannot–be measured are properly documented.*
    - *Manage 1.2: Treatment of documented AI risks is prioritized based on impact, likelihood, and available resources or methods.*
    - *Govern 4.2: Organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate, and use, and they communicate about the impacts more broadly.*