# FLI Response to OMB: Request for Comments on AI Governance, Innovation, and Risk Management

21st February 2024

**US Policy Team**

policy@futureoflife.org

# Request for Comments on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence Memorandum

## Organization

Future of Life Institute

## Point of Contact

Hamza Tariq Chaudhry, US Policy Specialist. hamza@futureoflife.org

We would like to thank the Office of Management and Budget (OMB) for the opportunity to provide comments on OMB–2023–0020, or the Memorandum on 'Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence' (hereafter referred to as 'the Memorandum'). The Future of Life Institute (FLI) has a long-standing tradition of work on AI governance to mitigate the risks and maximize the benefits of artificial intelligence. For the remainder of this Request for Comment (RfC) document, we provide a brief summary of our organization's work in this space, followed by substantive comments on the Memorandum. The 'substantive comments' section provides responses to the questions outlined in the RfC. The 'miscellaneous comments' section offers general comments outside the scope of the questions outlined in the Federal Register. We look forward to continuing this correspondence and being a resource for the OMB's efforts in this space in the months and years to come.

## About the Organization

The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence. Since its founding ten years ago, FLI has taken a leading role in advancing key disciplines such as AI governance, AI safety, and trustworthy and responsible AI, and is widely considered to be among the first civil society actors focused on these issues. FLI was responsible for convening the first major conference on AI safety in Puerto Rico in 2015, and for publishing the Asilomar AI principles, one of the earliest and most influential frameworks for the governance of artificial intelligence. FLI is the UN Secretary General's designated civil society organization for recommendations on the governance of AI and has played a central role in deliberations regarding the EU AI Act's treatment of risks from AI. FLI has also worked actively within the United States on legislation and executive directives concerning AI. Members of our team have contributed extensive feedback to the development of the NIST AI Risk Management Framework, testified at the Senate Insight Forums, participated in the UK AI Summit, and helped connect leading experts in the policy and technical domains to policymakers across the US government.

FLI's wide-ranging work on artificial intelligence can be found at futureoflife.org.

# Substantive Comments

## Comments in Response to Questions 5 and 6 from the Federal Register: Definitions of and best practices regarding safety-impacting and rights-impacting AI.[1]

The Memorandum establishes a clear minimum threshold for safety (Section 5,c,iv) that must be attained before agencies are allowed to use AI systems, applicable to both systems being used presently and those intended for use in the future.[2] The requirements for these agencies - which include impact assessments, real-world testing of AI systems before deployment, independent evaluations and periodic post-deployment testing - are a positive step towards minimizing the safety risks from government use of AI models.

We would, however, welcome further details for agencies on the periodic reviews that occur post-deployment to specify that these reviews would also include red-teaming and other auditing processes that make up portions of the pre-deployment review process. In addition, while we appreciate the inclusion of language prohibiting agencies from using AI systems in cases where 'the benefits do not meaningfully outweigh the risks,' we invite the OMB to support this language with quantitative examples, as risks may capture both probability and magnitude of harm, especially in the case of safety concerns. For instance, even if the probability of any given risk is found to be considerably lower than that of potential benefit, the magnitude of a risk (e.g., a bio-weapon attack) may be so high that it overrides the benefit despite being of low probability. Agencies should be required to establish, subject to public comment and external review, risk tolerances for activities for which use of AI systems is anticipated, including unacceptable risks to individuals, communities, and society that would disqualify the system from adoption. Establishing these thresholds prior to testing and adoption would help prevent drift in risk tolerance that could gradually rise to insufficient levels.

The Memorandum provides adequate definitions for two categories of potential harm posed by AI systems - safety-impacting AI systems and rights-impacting AI systems. FLI, which predominately focuses on AI safety, supports the broader definition of safety-impacting AI systems offered in the Memorandum, which captures a more expansive set of AI models and does not rely on technical thresholds. We believe this best positions the executing Agencies to exercise appropriate oversight over use of AI models. In addition, we are pleased to see that under the proposed definition, many models are presumed to be safety-impacting (Section 5,b). This is critical as it relieves relevant agencies of administrative burdens and time delays that would otherwise occur in evaluating each system with risk assessments, instead allowing them to devote more time and resources to setting up adequate guardrails. On the same token, we are pleased that additional risk assessments can be conducted to expand the scope of systems receiving due scrutiny.

Finally, when it comes to 'use of AI,' we support efforts to include cases in the Memorandum of procurement in addition to direct use (Section 5, d). However, the language of the Memorandum currently forwards guidance on procurement and contracts not as a set of requirements but as a set of recommendations. It is imperative that the OMB set up robust requirements for government purchasing of AI systems that mirror requirements on direct use, ensuring that procurement of AI systems includes consistent, robust evaluation to protect the safety and rights of the American public. This has the potential to minimize harm from government use of AI, and to inform best practices for the private sector, where most of that state-of-the-art models are created.

[1] 5. Are there use cases for presumed safety-impacting and rights-impacting AI (Section 5 (b)) that should be included, removed, or revised? If so, why? 6. Do the minimum practices identified for safety-impacting and rights-impacting AI set an appropriate baseline that is applicable across all agencies and all such uses of AI? How can the minimum practices be improved, recognizing that agencies will need to apply context-specific risk mitigations in addition to what is listed?

[2] We are particularly pleased to see that the scope of this Memorandum applies not just to use and application of AI systems in the future, but also those currently in use by relevant agencies.

## Comments in Response to Questions 1 and 2 from the Federal Register: Role of Chief AI Officer and the benefits and drawbacks of central AI governance body.[3]

We agree that effective oversight of AI adoption by government agencies should rely on AI governance bodies within each agency to coordinate and supervise AI procurement and use across the broad functions of the agency.This structure facilitates oversight and accountability  to ensure that minimum requirements as set out in the Memorandum are met by each agency writ large, while giving different offices within each agency the capability to exercise their mandate when it comes to specific use cases. In addition, we believe such a body can facilitate effective communication across different offices, bureaus and centers within the agency to ensure that poor communication feedback does not lead to under-reporting of use cases or use of AI that could lead to potential harm. Finally, we believe such a body would appropriately empower the Chief AI Officer (CAIO) to exercise their mandate as specified in the Memorandum.

However, we contend that this "hub and spoke" structure of a centralized AI governance body coordinating and overseeing domain-specific AI governance should be implemented on a whole-of-government level. In other words, we believe that just as there are benefits to having a new central body within each agency that helps enforce requirements laid out within the Memorandum, these bodies themselves would benefit from a single governance body that has representation and oversight across different agencies. This would facilitate interagency coordination, provide a central hub of expertise to advise agencies where appropriate, avoid costly redundancies in efforts by various agencies, and provide a body to inform and evaluate government AI adoption where domain-specific agency jurisdiction is not clear.

## Comments in Response to Question 8 from the Federal Register: Nature of information that should be publicly reported by agencies in use case inventories.[4]

While we welcome provisions within the Memorandum which require annual reporting of use cases of covered AI systems by the relevant agencies (Section 3, a), we are concerned that further elaboration is not provided by the OMB on the details of these use case inventories. We believe that the public should have access to information on the full results of the impact assessments, real-world testing, independent evaluations, and periodic human reviews, wherever possible. Where it is not possible to provide this information in full, we believe it is vital to provide redacted iterations of these documents upon the filing of a Freedom of Information Act (FOIA) request. Secondly, considering that there is some precedent of agencies neglecting to report all use cases in the past, we believe that the Memorandum would benefit from having explicit provisions to guard against under-reporting of use cases. This could, for instance, include guidance for Inspectors General to audit these use cases periodically within their respective agencies. Finally, while we recognize this as a positive first step towards creating transparency in use cases, we emphasize that this does not ensure sufficient accountability in and of itself, and will require further guidance and requirements on empowering the OMB and the CAIOs, and other relevant authorities, to take against violations of use case guidance set up in the Memorandum.

[3] 1. The composition of Federal agencies varies significantly in ways that will shape the way they approach governance. An overarching Federal policy must account for differences in an agency's size, organization, budget, mission, organic AI talent, and more. Are the roles, responsibilities, seniority, position, and reporting structures outlined for Chief AI Officers sufficiently flexible and achievable for the breadth of covered agencies? 2. What types of coordination mechanisms, either in the public or private sector, would be particularly effective for agencies to model in their establishment of an AI Governance Body? What are the benefits or drawbacks to having agencies establishing a new body to perform AI governance versus updating the scope of an existing group (for example, agency bodies focused on privacy, IT, or data)?

[4] 8. What kind of information should be made public about agencies' use of AI in their annual use case inventory?

# Miscellaneous Comments

## Comments on Scope

Section 2 (‘Scope’) explicitly exempts the intelligence community (‘covered agencies,’ Section 2, a) and cases where AI when it is used as a component of a national security system (‘applicability to national security systems,’ Section, c). As the Memorandum is intended to minimize the risks of government use of AI systems, we believe it is critical to establish robust requirements for the intelligence and defense communities, as these are likely to be the highest risk cases of government AI use with the greatest potential harm, and hence the most urgent need for scrutiny. Where it is within the remit of the OMB to set up requirements within these domains, we ask that they urgently do so.

## Comments on Definitions

We are pleased to see that Section 6 (‘Definitions’) outlines an expansive definition of “artificial intelligence” that is broader than the definition offered in the AI Executive Order. In addition, we support that the Memorandum’s description of AI systems encompasses all those across different ranges of autonomous behavior, technical parameters and human oversight. However, we believe that it is vital to ensure that the definition of AI employed in this section is treated as an ‘or’ definition as opposed to an ‘and’ definition. In other words, we believe that any system which fulfills any of these criteria should fall within the definitional scope of AI. For the same reason, we are concerned that the definition of ‘dual-use foundation models’ mirrors the definition included in the AI Executive Order, which offers an ‘and’ definition leading to very few models coming under definitional scope, and potentially excluding those which pose safety risks but do not meet other criteria.[5]

The Memorandum also employs the AI Executive Order definition for ‘red-teaming.’[6] While this definition outlines what red-teaming would cover, it does not provide any detail on how rigorous this red-teaming must be, and for what period within the lifecycle of the AI system. We support further clarification from the OMB in this regard to ensure that red-teaming as defined in guidance adequately tests models for safety harms for the duration of their procurement and use.

We endorse the OMB’s decision to establish a broad definition for what would count as ‘risks from the use of AI’ as well as the expansive definition of ‘safety-impacting AI.’ However, we recommend the addition of loss of control from use of AI systems to the considerable list of risk factors identified in the definition of ‘safety-impacting AI.’

## Comments on Distinguishing between Generative and Other AI

We believe that all advanced AI systems, whether they are generative or otherwise, should be subject to appropriate requirements to ensure safety. Hence, we are pleased to see that, in a slight divergence from the AI Executive Order, the Memorandum bases requirements on potential harms from AI and does not distinguish between generative AI and other AI systems.

[5] Section 3 of the AI Executive Order defines such as model in the following way: “dual-use foundation model” means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; *and* that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters.” (Emphasis added).

[6] Section 3 of the AI Executive Order defines red-teaming as: The term “AI red-teaming” means a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Artificial Intelligence red-teaming is most often performed by dedicated “red teams” that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system.