APOLLO
RESEARCH

Subject: Apollo Research Comment Regarding "NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)"
Reference: 88 FR 88368, Document Number 2023-28232

About Apollo Research

As increasingly capable artificial intelligence (AI) systems are developed and deployed across the global economy, it is crucial that the actions taken by these systems are safe, reliable and steerable. This is especially important for deployment in critical infrastructure such as in energy grid management, or high-risk uses such as clinical-decision support in medicine. Moreover, the social and economic dividends of AI can be best unlocked if its uses present a degree of risk from accidents, misuse or loss of control that citizens are willing to accept, and once we reach sufficient confidence in our methods of human oversight.

Apollo Research is an AI evaluations start-up established to meet this challenge by specializing in **evasion of human control through deception or obfuscation**, a key challenge highlighted in the White House Executive Order 14110 (EO) on AI.[1] We evaluate the most **advanced dual-use foundation models**, such as large language models (LLMs), which makes our work applicable across use cases and sectors where such AI systems could be implemented. Our current focus is conceptualizing and evaluating the capability for AI systems to deceive either the user or its designer, and the prerequisites to this capability such as situational awareness.

Apollo Research undertakes **behavioral evaluations, including red-teaming of dual-use foundation models,** some of which were showcased at the UK's AI Safety Summit hosted in Bletchley Park, November 2023. We also undertake mechanistic interpretability research with the goal of having a comprehensive understanding of what is driving AI systems' behaviors, capabilities and propensities.

We are committed to enabling safe and beneficial AI innovation. We believe that clear guidance and standards can: level the playing field for small businesses; derisk investments; help businesses plan and execute their market-access strategies; as well as avoid the costly need to retrofit products and services where regulations are introduced, as would likely occur once safety issues arise.

[1] For further details, see section 3(k) in the EO 14110.

Summary of key points

We are thankful for the opportunity to share some of our research and work through our response to this Request for Information (RFI). While the RFI has a broad remit, our submission focuses on our area of expertise which is evaluations of dual-use foundation models. We hasten to add that our submission is not a comprehensive reflection of the literature, but selectively highlights the work we considered most relevant for NIST to consider along with other submissions.

Below, we briefly summarize our key points in this submission.

(1) Defining and differentiating safety & security practices: auditing, evaluations, benchmarking, and red-teaming.

Accurate and functional definitions are foundational for a shared understanding among participants across the AI value chain, enabling NIST to successfully execute its assignments under the EO. We therefore offer the definitions we use and how they relate to each other.

(2) Processes and prerequisites for evaluating and managing dual-use foundation models AI systems. This section is divided into three parts:

> **(2.1) Processes for evaluating AI systems across the lifecycle.** Here we explain the process Apollo Research has followed when evaluating dual-use foundation models, methodological considerations and what we consider to be the current limitations of and promising paths forward for this nascent field.
>
> **(2.2) Prerequisites for evaluation** articulates resources and other requirements we consider absolutely essential for evaluation to take place, such as: defining the object of evaluation; degree of access to models; personnel and other resources.
>
> **(2.3) Prerequisites for an effective evaluation ecosystem** outlines our main observations on the wider governance structures that would enable more effective evaluations of dual-use foundation models. We also recommend that a government agency should monitor how effective evaluations are at preventing harms through proactive data collection, which in turn will support a better 'Science of Evaluations'.

(3) Leading responsible AI development through global technical standards and on the international stage. In this section, we recommend that NIST prioritizes alignment between the international and domestic standards on dual-use foundation models, how they are evaluated, and who undertakes or designs evaluations of them. We explain why this is the best 'gap' in global technical standards for NIST to tackle, and why building international agreements on this will be better for business and help maintain US competitiveness.

(1) Defining safety and security practices: auditing, evaluations, benchmarking, and red-teaming[2]

Accurate and functional definitions are foundational for a shared understanding among all actors across the AI value chain. As is natural for a fast-evolving technology, different interpretations are offered by stakeholders from different intellectual fields on key assurance practices for dual-use foundation models (e.g. Raji et al 2020; Shevlane et al 2023; Kinniment et al 2023). Nevertheless, there are few authoritative or widely recognised definitions.[3] We see this as an excellent opportunity for NIST to create standardized definitions of what constitutes an *evaluation*, *red-teaming* or *benchmarking* exercise, in order to enable greater confidence in these practices and more effective conversations between all participants across the AI value chain.

Below we contribute definitions of five terms we routinely use within our core work.

Apollo Research characterizes the following:

- **Evaluations ('evals')** are "the systematic measurement of properties in AI systems", including its:
    - capabilities (i.e. what a system can do), such as the capability to solve a specific coding problem;
    - propensities (i.e. the likelihood of the capability presenting across different scenarios or settings), such as tendency to be power-seeking, and;
    - alignment (i.e. the degree to which an AI system has a propensity to do things that are aligned with human intentions, or not), such as how consistently an AI system

[2] This content is most relevant to the following items in the RFI:

- "Roles that can or should be played by different AI actors for managing risks and harms of generative AI (e.g., the role of AI developers vs. deployers vs. end users)"
- "Definition, types, and design of test environments, scenarios, and tools for evaluating the capabilities, limitations, and safety of AI technologies"
- "Applicability of testing paradigms for AI system functionality, effectiveness, safety, and trustworthiness including security, and transparency, including paradigms for comparing AI systems against each other, baseline system performance, and existing practice, such as: model benchmarking and testing"
- "AI nomenclature and terminology" of relevance to global technical standards for AI development

[3] An important exception are definitions on auditing and conformity assessments which derive from existing general standards (e.g. ISO/IEC 17000:2020), which can then be developed into more specific standards for AI systems (e.g. ISO/IEC 42001:2023: ISO/IEC DIS 42006 - in draft). In light of the potential risks of dual-use foundation models outlined in the EO, we think there will likely be qualitative differences in the meaning of practices such as 'auditing' when applied to dual-use foundation models vis-a-vis other AI products being assessed for market access.

completes a task as intended by the AI developers training of or instructions given to it.
    - We consider red-teaming and benchmarking to be distinct sub-components of evaluations, both of which test different qualities of the model.
- **Red-teaming** is a type of evaluation that actively searches for specific capabilities while interacting with the specific AI system. It aims to demonstrate the existence of the capability but does not make a claim about the likelihood of the capability occurring in typical deployment. We consider red-teaming an essential component of evaluations; the more rigorous the searching for high-risk capabilities, the greater the potential for justifiable confidence in the AI system.
- **Benchmarking** is a type of evaluation that aims to identify the likelihood of an AI system behaving in a specific way on a certain range of inputs, typically to understand the likelihood of a behavior occurring under real-use conditions.
- **Auditing** refers to a holistic suite of practices that include evals, governance auditing, compliance auditing, and more. We consider evaluations to be a part of auditing.

For a more in-depth explanation, please refer to A Starter Guide for Evals — Apollo Research. We also recommend the UK AI Safety Institute's definition of evaluations, including the different properties (i.e. capabilities and risks) of an AI system which an evaluation tests (Department of Science, Innovation and Technology 2023; pages 8-10).

(2) Processes and prerequisites for evaluating and managing dual-use foundation models

(2.1) Processes for evaluating AI systems across the lifecycle[4]

In this section, we draw on our red-teaming of dual-use foundation models for capabilities of concern such as deception (Scheurer, Balesni & Hobbhahn, 2023) to *outline relevant steps in the evaluation process.* This is relevant for questions around designing and implementing evaluation or red-teaming exercises and the current availability of methods for measuring the properties and capabilities of AI systems.

[4] This content is most relevant to the following items in the RFI:
- "Definition, types, and design of test environments, scenarios, and tools for evaluating the capabilities, limitations, and safety of AI technologies"
- "Current red-teaming best practices for AI safety, including identifying threat models and associated limitations or harmful or dangerous capabilities"
- "Sequence of actions for AI red-teaming exercises and accompanying necessary documentation practices"
- "How to design AI red-teaming exercises for different types of model risks, including specific security risks (e.g., CBRN risks, etc.) and risks to individuals and society (e.g., discriminatory output, hallucinations, etc.)"

(i) Defining the threat model: To design effective evaluations, there needs to be a clear definition of the behavior of interest and the threat model for how it could lead to harm (which in turn helps generate a risk profile for the model). Otherwise, the evaluators risk measuring the wrong property and drawing incorrect conclusions. This includes engaging with relevant literature, other researchers, and subject experts from a range of fields, if required for the specific threat model and evaluation (e.g. for chemical, biological, radioactive or nuclear threats; CBRNs). Similarly, it is important to ensure the construct validity of the evaluations. In order to do so, conceptualisation and definition of the threat model and examples of how it might present (i.e. the behavior) should undergo external feedback - and ideally peer review - prior to implementation. An example of such conceptual work is available in our publication on 'Understanding strategic deception and deceptive alignment'.

(ii) Developing scenarios and environments for the evaluation: In essence, the goal of evaluators is to find a diverse set of individual evaluations that only share the behavior of interest and nothing else, to avoid measuring a proxy of the desired property. For example, violent language often carries negative sentiment. If we solely want to measure the effect of violent language during an evaluation, then we also have to include settings that exclude negative sentiment to avoid measuring a proxy. Scenarios and environments for evaluations can be designed in many different ways, e.g. by humans or (as is increasingly popular) with the help of other AI models (see e.g. Perez at al., 2022).

(iii) Robustness testing: This refers to the importance of frequently and extensively testing the robustness of the evaluation, to ensure that it is measuring and honing in on the specific behavior it is meant to identify. Experiments should be re-run multiple times to report on and estimate any variation in results. We recommend varying individual parts of the evaluation in a controlled setting to better understand their impact on the result. In LLM evaluations, for example, the exact wording of the prompt can have a surprisingly large effect on the final outcome (e.g. Liang et al., 2022; Mizrahi et al., 2023; Scalar et al., 2023). In this specific case, we would recommend averaging results over multiple prompts with similar meaning but different wording. In our work on deception (Scheurer et al, 2023), our robustness tests included varying the strength of prompts as well as safeguards, such as system prompts. We also tested whether the behavior could be triggered by specific cues, as could occur with injection attacks, jailbreaks, or by natural use.

(iv) Interpreting empirical work: During this phase, the results are put into the context of the threat model and their implications are interpreted. In our work on deception, we also reviewed how variations of the experiment affected the frequency of the behavior to infer what might be its core drivers. Given the complex nature of highly-capable AI systems, we especially recommend investigating potential confounders and alternative hypotheses that could explain the results.

For example, in our work on deception we gave the AI system access to a scratchpad for chain-of-thought reasoning. This had the extra benefit of giving us insights into rationale for the behavior provided by the AI system. But, we also considered the possibility that the access to the

scratchpad might change the text distribution in a way that induces the LLM to role-play a deceptive character (Shanahan et al., 2023; Shah et al., 2023), which would have confounded our results. Therefore, we ran an additional experiment to test the impact of having access to or not having access to a scratchpad on deceptive behavior and updated our interpretations accordingly.

(v) Reporting findings: In many cases, there is a specific group of people who either commissioned the evaluation, such as AI companies or governments, or are otherwise a special target group, such as the AI research community. It may make sense to inform a government agency or AI safety institute about noteworthy outcomes of specific evaluations by default. At the same time, in cases where it is unclear if the findings should be public, because of e.g. potential for misuse, it might make sense to discuss and plan publication with such an agency before proceeding. We discuss this further in section 2.3.

(vi) Peer review / external scrutiny: If the results of the evaluation are deemed safe to publish, it is typically beneficial to do so and to explicitly invite external scrutiny and peer review. Feedback from the wider community led us to rerun a variation of the deception experiment, improving the final paper with updates to the methodology and analysis of results.

(2.1.1) Limitations of current techniques, methods and measurements[5]

(i) Generalisability of methods: Given the current nascency of the field, it is as of yet unclear as to how general the insights from AI system evaluations can be treated. Since absence of evidence is not evidence of absence, it is hard to know whether a model does not have that property or whether there was a flaw in the experiment such that it failed to identify that property. We would need stronger evaluation methods, such as interpretability-based evaluations, to have greater confidence in the absence of particular model properties.

Furthermore, when we evaluate a model for one capability, such as persuasion or manipulation, it is unclear how much evidence that provides about a very related property, such as deception. This determines how fine-grained our evaluations should be to ensure large coverage of the relevant behavior. Additionally, a 'defense in depth' approach to evaluations, involving a variation of approaches to testing a given property, would unlock greater confidence in results produced.

---

[5] This content is most relevant to the following items in the RFI:

- "Current techniques and implementations, including their feasibility, validity, fitness for purpose, and scalability, for: Model validation and verification, including AI red-teaming"
- "Generalizability of standards and methods of evaluating AI over time, across sectors, and across use cases"
- "Internal and external review across the different stages of AI life cycle that are needed for effective AI red-teaming"
- "Limitations of red-teaming and additional practices that can fill identified gaps"

Ultimately, the lack of generalisability should be expected for such a young field. This necessitates a broader range of foundational work on the 'Science of Evaluations' to ensure that the field is on par with other high risk sectors. In the referenced paper, we put forward a number of open research questions on both the trustworthiness of results as well as measurement of the right property.

(ii) Robustness of benchmarks: We think that there are well-reasoned critiques of benchmarks, such as their vulnerability to being 'gamed' (Anderljung et al 2023) or the fact that they can be saturated quickly and lose their utility (Maslej et al 2023). There are ways to address some of these problems, such as by preventing AI developers from accessing benchmarks, developing dynamic benchmarks (see Kiela et al 2021) or simply developing a wide-range of evaluation experiments. In addition to these solutions, we propose that it is important to pair benchmark assessments with red-teaming in order for the evaluation of potential harms to be as thorough as possible.

(iii) Behavioral nature of evaluations: Currently, most evaluation protocols only look at input-output relations and treat the model as a black box. While this allows us to interpret and build theories about the mechanisms underlying that behavior, it is often hard to verify these theories empirically. We recommend evaluations are developed to include interpretability methods (as they improve), so that it is possible to investigate both the behavior and likely causal drivers based on model internals.[6]

(iv) Independent red-teaming: Previously, we proposed that red-teaming is essential for effective evaluation. Now, we outline the added-value of *independent red-teaming*, by which we mean it is undertaken by individuals who are not employed by or otherwise beholden to the company who built the AI system that is being scrutinized.

Firstly, the effectiveness of red-teaming is contingent on the developers of the AI system being blind to the methods used by the red-teamers. If that is not the case, they may unconsciously or consciously optimize their development work to 'pass' tests without necessarily taking commensurate steps to address and rectify the underlying risks. Red-teamers should be safeguarding the privacy of their methods, without those being compromised by social ties or interactions between red-teamers and developers, leading to e.g, accidental 'leakage' of methods.

Secondly, we would like to draw attention to the fact that there are noteworthy instances in other industries where a lack of independent safety assessment was linked to serious accidents. For example, formal investigations into the Boeing 737 MAX crashes identified issues with how significant assurance responsibilities were in effect delegated in-house to manufacturers (Office of Inspector General, US Department of Transportation, 2021). In light of this, an additional benefit of independent red-teaming is the *increased assurance it offers the public*.

[6] For a thorough discussion of this topic, we recommend Caspar et al 2024.

Finally, promoting independent red-teaming will likely grow the pool of red-teamers, leading to a greater variety and novelty of approaches which can offer 'defense in depth' and therefore better assurance for all.

(2.2) Prerequisites for evaluation[7]

In this section, we explain the dependencies for evaluation, such as resources or agreements between stakeholders, without which evaluations either cannot happen or their utility would be compromised.

(i) Define and agree upon the object of evaluation: We consider the appropriate object of an evaluation to be the AI system, including *all affordances that could realistically be made available* to it. By affordances we mean the environmental resources and opportunities for affecting the world that are, or could be, available to a dual-use foundation model and could thereby affect its capabilities and risk profile. For example, affordances might include post-training enhancements such as tools (e.g. web search) intended to augment the system, or the number and variety of people who have access to it, which can affect likelihood of misuse. We propose that this more holistic evaluation approach will better reflect real-world conditions that could lead to misuse or evasion of control, and therefore be more robust than a traditional approach. A more in depth discussion is provided in Sharkey et al (2023).

This recommendation represents a shift away from thinking about evaluating AI systems designed for specific use-cases. Instead, it is geared towards evaluating the range of capabilities and use cases a given AI system could realistically acquire through actions undertaken by either its developers and users. This approach would make evaluations more future-proof in light of technological progress and application areas. Moreover, this approach is especially appropriate for dual-use foundation models because users can adjust them for more specific tasks Davidson et al 2023).

(ii) Access to the AI system, and other resources: Different degrees of access are needed for different types and depths of evaluations. Below, we outline a selection.

---

[7] This content is most relevant to the following items in the RFI:

- "Use cases where AI red-teaming would be most beneficial for AI risk assessment and management"
- "Capabilities, limitations, risks, and harms that AI red-teaming can help identify considering possible dependencies such as degree of access to AI systems and relevant data"
- "Information sharing best practices for generative AI, including for how to share with external parties for the purpose of AI red-teaming while protecting intellectual property, privacy, and security of an AI system"
- "Guidance on the optimal composition of AI red teams including different backgrounds and varying levels of skill and expertise"
- "The appropriate unit of analysis for red teaming (models, systems, deployments, etc.)"

- For exploratory efforts and a subset of behavioral evaluations, inference access (e.g. through the API of a model provider) can often be sufficient (Apollo Research, 2024).
- To get a better understanding of how close a current model is to having a specific capability or propensity, evaluators can fine-tune it to have that property. For example, to test how effective the model could be at user manipulation, the evaluator could actively fine-tune the model to be more manipulative to get a better sense of the *potential rather than existing capabilities*. Furthermore, some evaluations explicitly require fine-tuning access. For example, an evaluator might want to test how easy it is to remove the guardrails of a model through fine-tuning.
- For many evaluations, it is beneficial to have access to multiple types of models regardless of the depth of access. For example, it can be helpful to have access to model families (i.e. models that are based on the same architecture but in different sizes) to better understand the scaling trends for specific properties and make better predictions about future models. Furthermore, to investigate specific negative properties, having access to guardrail-free models can be useful. Otherwise, it is harder to differentiate whether the model is not capable of that action at all, or whether its intact guardrails (which could be circumvented in future attacks) were effective at that point in time.

In addition to different levels of access, secure sharing platforms could facilitate evaluations. The aim of secure sharing platforms is to reduce the risk of IP theft and privacy problems while external evaluators investigate the system. The platform can be configured such that neither the evaluator can steal weights from the AI system owner, nor that the AI developer sees the evaluation techniques taking place.

(iii) Team composition: A model evaluations team requires technical expertise, and typically subject matter expertise if the threat model being investigated is highly contextual and therefore requires (such as for disinformation, CBRNs or other abuse of AI by adversaries). The technical expertise is necessary to set up the experiments and ensure they are run properly; without it, the wrong hyperparameters could be chosen or experiments cannot even be designed and implemented.

For high-context evaluations, subject matter expertise is critical to design the evaluation, put the results into context, and consider potential alternative explanations. If subject matter expertise is lacking, the evaluators run the risk of executing the wrong experiments or misinterpreting the results.

(2.3) Pre-requisites for an effective evaluation ecosystem[8]

Previously, we outlined a basic structure for undertaking an evaluation. In this section, we share our observations on how evaluation practices overall could be more effective. We also put forward recommendations as to how NIST could help develop a more robust evaluation ecosystem.

(i) Availability of methods, metrics and benchmarks for measuring functionality, capability and limitations of dual-use foundation models: The RFI requested additional information on the availability of methods, metrics and benchmarks for measuring functionality, capability and limitations of dual-use foundation models. We want to especially highlight the following:

- Several benchmarks or metrics have been developed to measure capabilities of dual-use foundation models, such as MMLU (Hendrycks et al, 2020) and HELM (Liang et al, 2022).
- We also welcome further development of benchmarks relevant for different threat models; for example, the SAD-influence situational awareness benchmark for LLMs (Laine et al 2023) which is relevant for evasion of control risks.
- Benchmarks are being developed to test abilities that are fundamental to intelligence such as ConceptARC (Moskvichev, 2023), which assesses abstract reasoning.

[8] This content is most relevant to the following items in the RFI:

- "Roles that can or should be played by different AI actors for managing risks and harms of generative AI (e.g., the role of AI developers vs. deployers vs. end users);"
- "Governance policies and technical requirements for tracing and disclosing errors, incidents, or negative impacts"
- "The possibility for checks and controls before applications are presented forward for public consumption."
- "Availability of, gap analysis of, and proposals for metrics, benchmarks, protocols, and methods for measuring AI systems' functionality, capabilities, limitations, safety, security, privacy, effectiveness, suitability, equity, and trustworthiness, including:
    - Negative effects of system interaction and tool use, including from the capacity to control physical systems or from reliability issues with such capacity or other limitations
    - Risks arising from AI value chains in which one developer further refines a model developed by another, especially in safety- and rights-affecting systems
    - Impacts to individuals and society; including both positive and negative impacts on safety and rights."
- "Internal and external review across the different stages of AI life cycle that are needed for effective AI red-teaming"
- "Sequence of actions for AI red-teaming exercises and accompanying necessary documentation practices"
- "How AI red-teaming can complement other risk identification and evaluation techniques for AI models"

Despite a range of different evaluation methods currently used to understand the capabilities of dual-use foundation models, the field is overall still nascent and suffers from generalisability issues (see section 2.1.1). We therefore advocate for investment in a 'Science of Evaluations' to develop methods, metrics and measurements that are robust, reliable and reproducible. We also think that NIST can play a valuable role in this, as well as in codifying methods for an effective evaluation as the field is maturing.

(ii) Documentation: Clear guidance on the reporting of research and the standardization of documentation is central to raising the research integrity in high-risk fields. For example, the EQUATOR network produces a range of reporting guidelines for research in healthcare which can be adjusted as technologies or methodologies evolve. For the case of AI model evaluations, we suggest that similar guidelines on reporting could enable a better comparison of research results, and help non-experts critically appraise the quality of evidence that they are presented with. This could overall lead to a beneficial impact on public safety and raise citizen trust while harnessing innovation. We suggest that NIST is well placed to produce guidance on the minimum reporting requirements for AI evaluations.

In addition to this, it could be particularly helpful to encourage the publication of both methodologies and results in depth, to enable peer-learning. In-depth methodological publications by Model Evaluation and Threat Research (e.g. Kinniment et al 2023) have contributed significantly to the development of the field, as have their contributions to the GPT-4 model card (Open AI, 2023). These, and other evaluation reports we have cited, may be useful templates from which to begin developing reporting guidelines.

At the same time, there are potential trade-offs between research transparency and the prevention of proliferation of: a) results which could be exploited by malicious actors, or b) information on evaluation methodologies which could enable developers to 'game' evaluations. NIST may wish to consider developing guidance that would lay out relevant considerations for all actors in the evaluation ecosystem to navigate such trade-offs, including which organizations and / or subject-matter experts should be involved in the final decision-making process.

(iii) Checks and controls prior to public deployment: Once an AI system is deployed in public, it is harder to mitigate any shortcomings or threats that may have gone undetected. We therefore suggest that pre-public deployment checks are vital to mitigate societal harms. We outline a number of recommendations for pre-public deployment checks that AI developers and AI systems should undergo before deployment in Sharkey et al. (2023; in particular sections 2.3.6 and 2.37). For example, we propose that content moderation filters for the inputs and outputs of AI systems should be red-teamed and benchmarked against other systems to ensure they are technically adequate to prevent malicious and dangerous use. We also propose that developers should implement Know-Your-Customer ('KYC') constraints to determine which members of the public should be given which degree of access to which AI systems. Constraints such as these would benefit from independent, iterative third-party assessment of their adequacy.

(iv) Roles that should be played by different AI actors for managing risk and harms (including of generative AI): Currently, most independent AI evaluations occur just prior to deployment (e.g. OpenAI 2023; Ganguli et al 2023). This means that they occur at the end of the product development lifecycle. While this is still valuable, upstream decisions affect the capabilities and risk factors of AI systems (Sharkey et al 2023, section 2.3), such as choice of training data and the training-experiment design. This means that there is a broader opportunity space for risk management by independent third parties than is currently exploited.

We propose a range of *investigations that independent evaluators can perform across the AI life cycle* to improve risk identification and management, such as assessing the training data contents for potentially dangerous content, or risk-assessing different components of the training-experiment design. We comprehensively outline these and more investigations in Theories of Change for AI Auditing.

(v) Interaction between AI systems and available affordances: We suggest that independent evaluators can play an important role when it comes to reviewing, risk assessing and re-evaluating an AI system as its available affordances change. Available affordances such as system interaction and tool use can contribute to a change in risk profile and necessitate a review of previous risk assessments and evaluations (Sharkey et al 2023, section 2.3.1). Depending on the change, this may also necessitate review from additional relevant subject-matter experts.

An iterative evaluation process such as the one described will also contribute to shorter feedback loops. This will allow evaluators and other relevant parties to improve and adapt their methodologies swiftly in response to real-world evidence, minimizing significant upheaval for AI applications across services and enabling a swifter integration of the technology in all sectors.

(vi) Assessing the quality, accuracy and helpfulness of evaluations: The AI evaluations field, particularly with an eye to dual-use foundation models, is nascent. Yet, the risks from these systems have the potential to lead to severe consequences, and so necessitate very high confidence in the evaluation methods applied.

One recent investigation into lessons from existing assurance regimes looked at how the US Food and Drug Administration monitors medicines and medical devices, and proposed that a framework such as this could be used to collect *comprehensive data on real-time adverse events and near-misses* from AI systems to enable proactive safety monitoring (Ada Lovelace Institute, 2023). We suggest that such post-market surveillance methods provide an excellent blueprint for how the ecosystem can:

> a) identify where events such as harms or near misses are occurring, in order to;
> b) take protective action swiftly, including ceasing access to an AI system while investigations are underway, and;
> c) use this data to *review the effectiveness of the evaluation processes* to which the AI system was subjected.

Instances where risks were not identified by evaluations can and should contribute to thorough reviews of existing processes, including into fitness of the methodologies employed as well as fitness of the evaluator. For example, in the UK the Independent Medicines and Medical Devices Safety Review successfully scrutinized and made recommendations to the regulator for medicines and medical devices to improve its processes. We especially note that such reviews are typically reactively instigated in response to major safety or adverse events, *long after the harms have occurred* of, for example, adverse medical events, or false accusations of fraud or false accounting.

This is why we think that proactive and authoritative scrutiny is important to prevent AI harms proliferating due to how rapidly AI systems can be scaled up. We therefore suggest that a *government agency is designated as the responsible authority* for proactively collecting data on adverse events and near misses in order to monitor the effectiveness of different evaluations and, indeed, evaluators. This would close the feedback loop between evaluations and empirical data. We expect that NIST could be a natural home for this authority, due to its leadership on evaluations and intricate connection with the US AI Safety Institute.

(3) Leading responsible AI development through global technical standards and on the international stage[9]

In this section, we provide recommendations to advance safe AI development through engagement with global AI standards efforts. We believe that international leadership is needed to raise standards to a sufficient level in order to preempt and manage risks that could lead to global consequences. We recommend that the US should harness this opportunity to develop impactful standards for dual-use foundation models.

(i) Prioritize standard development for dual-use foundation models: As a first priority, we recommend that NIST should lead on the development of global technical standards for dual-use foundation models. This leadership should build upon existing guidance in the NIST AI Risk Management Framework (NIST AI RMF) and associated profiles such as the Berkeley General-Purpose AI Systems profile. In the interest of timeliness, these should start as internationally recognised voluntary standards and be reviewed iteratively (e.g. annually) to

[9] The content in this section is relevant to items in section 3 of the RFI, in particular:

- "Examples and typologies of AI systems for which standards would be particularly impactful (e.g., because they are especially likely to be deployed or distributed across jurisdictional lines, or to need special governance practices)"
- "Suggestions for AI-related standards development activities, including existing processes to contribute to and gaps in the current standards landscape that could be addressed, and including with reference to particular impacts of AI"
- "Potential mechanisms, venues, and partners for promoting international collaboration, coordination, and information sharing on standards development"
- "Potential implications of standards for competition and international trade"

ensure they reflect scientific developments. The goal should be to align internationally to one superseding set of standards and to incentivise adherence to them through access across international markets.

The prioritization we propose is both pressing and promising because:

- **The harms arising from dual-use foundation models transcend jurisdictional boundaries.** This is true of current harms supported by these systems (e.g. cyber-attacks, such as spear-phishing) and is equally likely the case for potential future harms, such as evasion of human control.
- **Dual-use foundation models currently lack formal standards**. As highly capable foundation models have only been released to the public since 2022, naturally the standards community's work has been more relevant for earlier AI systems with a narrower range of capabilities. We therefore see a gap in standards for advanced dual-use foundation models whose full capabilities and therefore risk profile emerges: a) in response to inputs (i.e. prompts); as well as b) affordances made available to them (e.g. scale and extent of deployment; post-training enhancements; Davidson et al 2023). Risk stratification of models will be important to ensure these standards are not unduly burdensome for models that would pose a low risk, and we recommend that available affordances are inputs for risk stratification (see 2.2.i).
- **Consistent, shared standards can give companies developing dual-use foundation models confidence that the due diligence they undertake will be advantageous on international markets**. While voluntary guidance is available, there are currently few incentives (such as access to government procurement frameworks, or a presumption of conformity with regulations) to follow such guidance. Without this, AI companies may question whether the costs of following guidance or adherence to standards outweigh the benefits. Such incentives are not only necessary but are also better enacted on the international level to promote a 'race to the top'. International consistency in the safety requirements that AI companies must meet reduces the opportunity to exploit citizens and customers in under-regulated markets.
- **Frameworks developed for today's dual-use foundation models will likely accelerate the development of standards for systems that will become more mainstream in the near future**. We reviewed voluntary guidance developed for large language models (LLMs), such as the UK government's Emerging Processes and the Berkeley General-Purpose AI Systems profile (which builds on the NIST AI RMF), and expect that many practices recommended here could offer a firm base to build on for, e.g., multi-modal models (MMM). These would include: systems for monitoring behavior; data input controls and audits; and security controls such as securing model weights.

(ii) Leverage standards to enable consistency in both the evaluations of dual-use foundation models, as well as the due diligence that evaluators go through: Evaluations are becoming increasingly prominent risk management levers, both within AI companies and nation states. We

therefore strongly encourage NIST to ensure that they remain effective levers by developing agreement, guidance and eventually standards on both:

- evaluations of advanced AI systems and;
- assurance requirements *for evaluators*, i.e. organizations who either undertake or design evaluations.

We propose these because:

- **Current standards for assurance or risk management of AI systems will not provide sufficient assurance for dual-use foundation models.** We commend US leadership of important AI standards committees, such as ISO/IEC JTC 1/SC 42, and work to standardize risk management practices for AI systems (e.g. ISO/IEC DIS 42006). At the same time, dual-use foundation models which push the frontier of the technology will require evaluations that search for both known and unknown risks. Assurance paradigms which focus on identifying and managing known risks, such as quality management systems, might not offer sufficient thoroughness.
- **Without independent guidance now, the standard for evaluations could be set too low**. Conversely, we think NIST and international partners can enhance the quality of evaluation through agreeing what constitutes an evaluation (as per section 2.1), what different actors must do to enable an evaluation to take place (as per section 2.2) and identifying gaps in the evaluation ecosystem which can be subsequently addressed through the standards process (as per section 2.3).
- **Assurance requirements for evaluators would bring this field in line with practices in other sectors and jurisdictions**. For example, notified bodies who approve high-risk products for the EU market, such as medical devices, must have their competence verified at regular intervals by an accreditation organization which is typically a government arms-length body (e.g. Irish National Accreditation Board). In a future where we have a scaled-up evaluation ecosystem without accreditation, AI companies and citizens have more reason to be skeptical of the results of evaluations and of the trustworthiness of the evaluators.

While this is substantive work, we think it would be most helpful to start building agreement and guidance on:

- The goals and object of an evaluation;
- Roles and responsibilities for all actors in the evaluation pipeline, including how the work of non-state evaluators is quality assured by national AI safety institutes;
- How access to AI systems is granted for evaluation, and the degree of access necessary;
- Responsibilities of all evaluators to protect and secure the AI systems that they are evaluating;
- How evaluation processes and results are documented and shared, such as through use of model cards (Mitchell et al, 2019).

(iii) Leverage existing fora for collaboration across nations, organizations and standards-bodies: International alignment of standards could offer significant economic benefit. Above, we recommended various domains in which standards should be prioritized and explained our rationale. Below, we conclude with highlighting a number of opportunities for NIST to collaborate on relevant international efforts. In particular, we suggest engaging with:

- The US-EU Trade and Technology Council (TTC); in particular through the Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management
- Continuing engagement with the international community as outlined in the Bletchley Declaration, including a commitment to appropriate collaboration on developing evaluation metrics, tools for safety testing, and relevant public sector capability and scientific research.
- The G7's Hiroshima process, building upon the Guiding Principles and Code of Conduct for organizations developing advanced AI systems
- The UK AI Safety Institute, and policy teams leading development of Emerging Processes for Frontier AI safety
- CEN-CENELEC Joint Technical Committee 21 'Artificial Intelligence', which is tasked with developing harmonized standards for the EU AI Act which will likely apply to dual-use advanced AI systems

We recommend that NIST, through its leadership of the US AI Safety Institute, develops norms for information sharing and collaborative, proactive risk management with like-minded countries and their national safety institutes. As a first step this should consider how *information on accidents, near misses or emerging capabilities can be shared*.

Conclusion

Our response elaborated on some of the existing strengths of the field, as well as gaps which present unique opportunities for NIST to improve the current quality of evaluations for dual-use foundation models. We provided a high-level overview of a number of steps necessary to conduct evaluations and hope to have shed light on the details and shortcomings of current processes. Iterative development of guidance and standards will present challenges due to how rapidly the frontier of AI shifts, but this also carries the opportunity for NIST to set the right course for AI evaluations early and to enable a quicker implementation of safe AI applications across multiple sectors.

Once again, we applaud NIST for its efforts and the opportunity to provide feedback. We remain available for further discussion and support at your request.

Contact us at: **governance@apolloresearch.ai**