Response to the RFI related to NIST's assignments under the Executive Order Concerning AI

Jonas Schuett
Research Fellow
Centre for the Governance of AI
jonas.schuett@governance.ai

Leonie Koessler
Research Scholar
Centre for the Governance of AI
leonie.koessler@governance.ai

Markus Anderljung
Head of Policy
Centre for the Governance of AI
markus.anderljung@governance.ai

February 2024

We welcome the opportunity to respond to the Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning AI. We offer the following submission for your consideration and look forward to future opportunities to provide additional input. Please note that our comments focus on "dual-use foundation models" as defined in Sec. 3(k) of the E.O. 14110.

About GovAI

The Centre for the Governance of AI (GovAI) is a nonprofit based in Oxford, UK. It was founded in 2018 as part of the Future of Humanity Institute (FHI) at the University of Oxford, before becoming an independent research organization in 2021. GovAI's mission is to build a global research community, dedicated to helping humanity navigate the transition to a world with advanced AI.

About the authors

- **Jonas Schuett** is a Research Fellow at GovAI. His research focuses on the regulation and governance of frontier AI models, with a special focus on risk management. Before joining GovAI, he advised the UK government on AI regulation and was part of Google DeepMind's Public Policy Team. He has a background in law.
- **Leonie Koessler** is a Research Scholar at GovAI. Her research focuses on frontier AI regulation and risk management, in particular risk assessment and technical standards. Before joining GovAI, she worked for the German government. She holds a Master of Laws (LL.M.) from King's College London.
- **Markus Anderljung** is Head of Policy at GovAI, an Adjunct Fellow at the Center for a New American Security (CNAS), and a member of the OECD Expert Group on AI Futures. His research focuses on the regulation and governance of frontier AI models. He was previously seconded to the UK Cabinet Office as a Senior Policy Specialist.

The views expressed in this submission are those of the authors and do not represent the views of GovAI.

Summary

1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

NIST AI RMF companion resource for generative AI, Sec. 4.1(a)(i)(A)

In general, we recommend that NIST:

- Draw from the UK Government's policy paper "Emerging Practices for Frontier AI Safety".
- Ensure that risk management burdens are proportionate to a system's impact.

Developers of dual-use foundation models should be encouraged to adopt a range of practices, including:

- Developing and publishing **comprehensive safety policies** that explain how they will avoid creating unacceptable risks.
- Developing **risk taxonomies** and **threat models** of the most severe risks.
- Conducting **model evaluations** and **red-teaming exercises** throughout the development lifecycle.
- Proactively identifying **specific evaluation results** that—in the absence of further safeguards—would indicate that a model poses an unacceptable risk.
- Producing (quantitative or semi-quantitative) **risk estimates** when making particularly high-stakes decisions, especially model release decisions.
- Combining **rules-based and risk-based approaches** to making decisions.
- Engaging in continuous **post-deployment monitoring** of models for signs of misuse, harm, and unexpected capabilities.
- Reporting **safety incidents** to competent authorities.

Guidance and benchmarks for evaluating and auditing AI capabilities, Sec. 4.1(a)(i)(C) and Guidelines for conducting AI red-teaming tests, Sec. 4.1(a)(ii)

- Developers should not only conduct model evaluations and red-teaming tests, but also aim to **advance the science**.
- Additionally, they should subject their models to **external scrutiny**.

2. Reducing the Risk of Synthetic Content

Report on standards, tools, methods, and practices related to synthetic content, Sec. 4.5(a)

- Developers should **develop and distribute tools** and methods to identify (or otherwise address risks from) synthetic content.

3. Advance Responsible Global Technical Standards for AI Development

Plan for global engagement on promoting and developing AI consensus standards, cooperation, and coordination, Sec. 11(b)

We recommend that NIST:

- Engage with the increasing number of AI Safety Institutes globally.
- Coordinate with European standard-setting processes.
- Participate in international standard-setting processes.

1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

NIST AI RMF companion resource for generative AI

Sec. 4.1(a)(i)(A) of the E.O. 14110 directs NIST to develop a companion resource to the NIST AI Risk Management Framework (AI RMF)[1] for generative AI. Below, we provide general comments, and make recommendations for each of the four functions defined in the AI RMF: Govern, Map, Measure, and Manage.

General comments:

- **Draw from the UK Government's policy paper "Emerging Practices for Frontier AI Safety".** Ahead of the Bletchley AI Safety Summit, the UK Department for Science, Innovation & Technology (DSIT) published a policy paper in which it sets out nine emerging practices for the safe development and deployment of frontier AI models.[2] NIST should take these practices into account when developing its companion resource for generative AI.

- **Ensure that risk management burdens are proportionate to a system's impact.** Dual-use foundation models might pose severe risks to society. They might have dangerous capabilities[3] that could be misused by malicious actors,[4] or they may have safety, bias, or privacy issues that lead to unintended harm.[5] For example, cybercriminals or terrorists might use them to conduct large-scale cyber attacks[6] or develop biological weapons.[7] The developers of dual-use foundation models should therefore take extensive measures to reduce the associated risks to an acceptable level. However, the same measures might be disproportionate for developers of less risky models. The companion resource should be sensitive to this distinction. A good proxy for a model's impact—though certainly not a perfect one—is the amount of computational resources ("compute") used to train it.[8] Another important proxy is the number of people who use or are affected by the model's outputs. For example, a model should receive greater scrutiny for bias if it may be used to inform very large numbers of decisions.

- **Address key governance challenges of dual-use foundation models.** The governance of dual-use foundation models poses distinct challenges, particularly in the context of threats to public safety. In our recent paper "Frontier AI regulation",[9] we discuss three key challenges: (1) models may possess unexpected and difficult-to-detect dangerous capabilities; (2) models deployed for broad use can be difficult to reliably

[1] NIST, Artificial intelligence risk management framework (AI RMF 1.0), 2023.
[2] DSIT, Emerging processes for frontier AI safety, 2023.
[3] Shevlane et al., Model evaluation for extreme risks, 2023.
[4] Anderljung & Hazell, Protecting society from AI misuse: When are restrictions on capabilities warranted?, 2023; Brundage et al., The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 2018.
[5] Weidinger et al., Ethical and social risks of harm from language models, 2021.
[6] Kaloudi & Li, The AI-based cyber threat landscape: A survey, 2020; Guembe et al., The emerging threat of AI-driven cyber attacks: A review,2022; Mirsky et al., The threat of offensive AI to organizations, 2023.
[7] Sandbrink, Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tool, 2023; Soice et al., Can large language models democratize access to dual-use biotechnology?, 2023; Urbina et al., Dual use of artificial intelligence-powered drug discovery, 2022; Mouton et al., The operational risks of AI in large-scale biological attacks, 2023; Nelson & Rose, Examining risks at the intersection of AI and bio, 2023.
[8] Sastry et. al. Computing power and AI governance, forthcoming; see also Anderljung et al., Frontier AI regulation: Managing emerging risks to public safety, 2023.
[9] Anderljung et al., Frontier AI regulation: Managing emerging risks to public safety, 2023.

control and to prevent from being used to cause harm; and (3) models may proliferate rapidly, enabling circumvention of safeguards. The companion resource should address these challenges.

Govern:

- **Developers of dual-use foundation models should have a safety policy aimed at reducing severe risks from their models.** Some developers have already published such policies. This includes Anthropic's Responsible Scaling Policy (RSP)[10] and OpenAI's Preparedness Framework (Beta).[11] Although the current versions of these policies are insufficient, they are an important step in the right direction. All developers of dual-use foundation models should have similar policies. The companion resource for generative AI should recommend creating such policies and provide further guidance.

- **Developers of dual-use foundation models should publish their safety policies.** Academic researchers, civil society organizations, the public, and regulators need to be able to scrutinize these policies and hold developers accountable.[12] NIST should therefore recommend that developers of dual-use foundation models publish their safety policies.

- **Developers of dual-use foundation models should only proceed with high-stakes development and deployment decisions if they have followed state-of-the-art safety practices.** Although safety practices are still emerging,[13] there are concrete actions developers of dual-use foundation models should take to reduce risks–ranging from risk assessment and risk management processes to information security practices. Many such practices are already included in the AI RMF[14] and more should be added in the companion resource for generative AI.

- **High-stakes development and deployment decisions should adhere to pre-defined evaluation-based decision-rules.** Existing safety policies describe potential results from model evaluations and specify safeguards that would keep risks to an acceptable level. Concretely, OpenAI's Preparedness Framework (Beta) defines risk levels based on whether a model is able to produce certain outputs, and specifies that models will only be deployed if their risk is "medium" or below.[15] Anthopic's Responsible Scaling Policy (RSP) describes AI Safety Levels (ASLs) based on model evaluation results, and only allows models to be developed and deployed if the corresponding safeguards are met.[16] Evaluation-based decision-rules along these lines are advisable in the development of the most impactful models since they set clear red lines and are directly tied to factors developers have control over.

- **High-stakes development and deployment decisions should be informed by attempts to estimate the level of risk imposed by the decision.** Evaluation-based decision-rules have many advantages, but it is difficult to judge whether adhering to them reduces risk to an acceptable level. Further, they fail to comprehensively consider the risk landscape. Finally, they may not allow developers of dual-use foundation models to prioritize between different ways to allocate scarce risk management resources. As such, evaluation-based decision-rules should be supplemented with attempts to explicitly assess the level of risk imposed by a given development or deployment decision. This should include estimates of the likelihood and impact of certain risk events, such as aiding a biological weapons attack or a large-scale cyberattack. Key risk factors should also be estimated, such as the chance that a model is stolen or leaked. Producing these estimates is

[10] Anthropic, Responsible scaling policy, 2023.
[11] OpenAI, Preparedness framework (Beta), 2023.
[12] Anderljung et al., Towards publicly accountable frontier LLMs: Building an external scrutiny ecosystem under the ASPIRE framework, 2023.
[13] DSIT, Emerging processes for frontier AI safety, 2023.
[14] NIST, Artificial intelligence risk management framework (AI RMF 1.0), 2023.
[15] OpenAI, Preparedness framework (Beta), 2023.
[16] Anthropic, Responsible scaling policy, 2023.

challenging and methods for doing so are still under development. Nonetheless, they are likely to be informative and the techniques will develop as more resources are invested in this area of study. Estimates of risk should directly inform decisions of whether to develop or deploy a specific model, via pre-defined risk thresholds. For example, a developer can say that they would only deploy a model if it increases the chance of a biological weapons attack on U.S. soil by no more than .1%. Risk estimation can also be used to assess the appropriateness of rules, including evaluation-based decision-rules.[17]

- **Developers of dual-use foundation models should use a combination of rules-based and risk-based approaches.** Having high-stakes development and deployment decisions informed by both whether pre-defined rules are followed and whether estimated risk is below pre-defined thresholds is common across safety-critical regulatory domains, including aviation, pharmaceuticals, and nuclear power. In the latter, operators must follow a long list of concrete regulations in addition to keeping the risk of radiation leakage below 10^{-5} per year, as assessed via probabilistic risk assessments. NIST should recommend that developers of dual-use foundation models use a combination of approaches to keep risk to an acceptable level, namely following pre-defined safety practices, such as adhering to evaluation-based decision-rules, and comparing risk estimates to pre-defined risk thresholds. We illustrate this in the figure below.

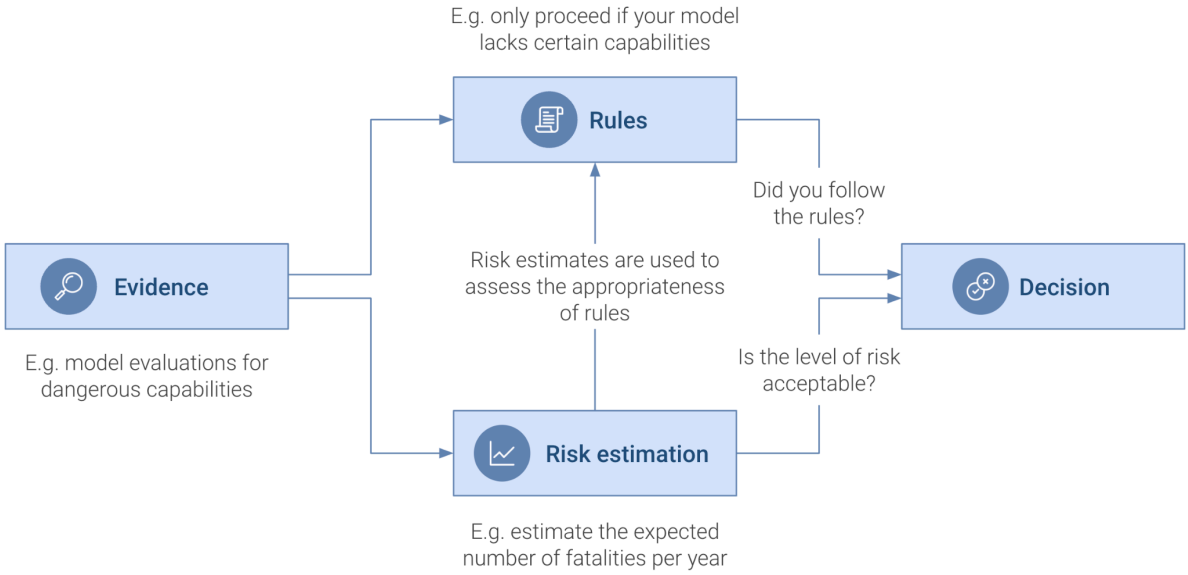

Figure 1: Illustration of how rules and risk estimation can jointly inform high-stakes development and deployment decisions[18]

- **Developers of dual-use foundation models should continuously monitor deployed models for signs of misuse and unexpected capabilities.** The companion resource should provide further guidance on specific post-deployment monitoring measures developers of dual-use foundation models could take.[19]
- **Developers of dual-use foundation models should report safety incidents to competent authorities.** This should include information about dangerous model capabilities and attempts to circumvent safeguards, where appropriate. Several AI companies already made voluntary commitments along these lines, but they

[17] This is analogous to the Nuclear Regulatory Commission's concept of "Risk-informed regulation".
[18] This figure is based on upcoming work by the authors.
[19] For more information, see Anderljung et al., Frontier AI regulation: Managing emerging risks to public safety, 2023.

need further specification.[20] The companion resource for generative AI should contain further guidance on incident reporting. NIST might draw from an information-sharing regime for developers of foundation models and the UK Office for AI proposed by GovAI researchers.[21]

- **Developers of dual-use foundation models should follow best practices in risk governance.** The companion resource for generative AI should recommend specific risk governance practices. This might include setting up a board risk committee, appointing a chief risk officer (CRO) or equivalent, implementing a version of the Three Lines Model[22] or Combined Assurance Framework, establishing a central risk function, and an internal audit function.[23]

Map:

- **Developers of dual-use foundation models should create risk taxonomies.** Risk taxonomies structure the overall risk universe and can serve to identify previously unknown risks. They provide the basis for comprehensive risk management efforts, helping to ensure that no important risks are overlooked.[24] Risk taxonomies can be developed for risks of societal harm,[25] or precursor events, such as cyber attacks that may target developers of dual-use foundation models[26] or adversarial attacks that may compromise the safety of AI systems.[27] The companion resource should recommend creating, using, and continuously updating risk taxonomies. It could also provide a template risk taxonomy, maybe even an example, that developers can use as a starting point.

- **Developers of dual-use foundation models should create detailed threat models of the most severe risks.** These should at least include risks related to CBRN weapons, cyber attacks, and the evasion of human control or oversight. Threat models should probably also distinguish between different actors that may drive the risk (e.g. lone wolves, terrorist groups, cybercriminal groups, nation states, or autonomous AI agents). The companion resource should highlight the importance of threat modeling and provide further guidance.

Measure:

- **Developers of dual-use foundation models should conduct model evaluations** for dangerous capabilities, in addition to evaluations for bias and other sources of social harm**.** The companion resource should recommend conducting model evaluations at key checkpoints during the development and before deploying a model. It should highlight the importance of evaluating dangerous model capabilities.[28] This might include, for example, the ability to deceive, persuade, and manipulate people,[29] find and exploit cyber vulnerabilities,[30]

[20] The White House, Voluntary AI commitments, 2023.
[21] Mulani & Whitlestone, Proposing a foundation model information-sharing regime for the UK, 2023.
[22] Schuett, Three lines of defense against risks from AI, 2023.
[23] Schuett, AGI labs need an internal audit function, 2023.
[24] Koessler & Schuett, Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries, 2023.
[25] E.g. Bommasani et al., On the opportunities and risks of foundation models, 2022; Critch & Russell, TASRA: A taxonomy and analysis of societal-scale risks from AI, 2023; Hendrycks et al., An overview of catastrophic AI risks, 2023; Shelby et al., Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2023; Weidinger et al., Taxonomy of risks posed by language models, 2022.
[26] Nevo et al., Securing artificial intelligence model weights, 2023.
[27] Vassilev et al., Adversarial machine learning, 2024.
[28] Shevlane et al., Model evaluation for extreme risks, 2023.
[29] Park et al., AI deception: A survey of examples, risks, and potential solutions, 2023; Hagendorf, Deception abilities emerged in large language models, 2023.
[30] Mirsky et al., The threat of offensive AI to organizations, 2021; Lohn & Jackson, Will AI make cyber swords or shields?, 2022.

provide instructions for the development of biological weapons,[31] or create copies of themselves and acquire resources.[32] Pre-deployment evaluations should also attempt to uncover biases, privacy-related issues, or other indications of social risks.[33]

- **Developers of dual-use foundation models should conduct red-teaming tests with a focus on dangerous capabilities.** The companion resource should recommend conducting red-teaming tests at several checkpoints throughout the lifecycle of a model. The focus should be on dangerous model capabilities, since red-teaming (relative to e.g. benchmarking methods) is currently particularly well-suited to the identification of dangerous capabilities. Red-teamers should include people internal and external to the organization with diverse backgrounds and expertise in the dual-use domains they are examining (e.g. cybersecurity, and biological weapons).[34] Red-teamers should be given sufficient access to the models for the respective tests they are conducting.[35]

- **Developers of dual-use foundation models should use quantitative or semi-quantitative approaches to estimate the likelihood and impact of relevant risks.** The companion resource should provide additional guidance and suggest concrete approaches for risk estimation. This might include a semi-quantitative approach similar to the approach used in the UK National Risk Register,[36] which uses likelihood and impact ranges. But it could also include a more sophisticated quantitative approach, for example, using the Delphi method to elicit expert estimates.

Manage:

- **Developers of dual-use foundation models should take adequate mitigation measures.** Developers of dual-use foundation models should take mitigation measures based on their risk assessments. These mitigation measures can be geared at what model capabilities to develop in the first place, how to deploy models, and who to grant access to models.[37] Common technical mitigation measures include fine-tuning through reinforcement learning from human feedback (RLHF)[38] or reinforcement learning from AI feedback (RLAIF),[39] as well as input and output filters.[40] Standard non-technical mitigation measures include employee

[31] Urbina et al., Dual use of artificial intelligence-powered drug discovery, 2022; Sandbrink, Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools, 2023; Soice et al., Can large language models democratize access to dual-use biotechnology?, 2023; Mouton, Lucas, & Guest, The operational risks of AI in large-scale biological attacks, 2023; Nelson & Rose, Examining risks at the intersection of AI and bio, 2023.
[32] Carlsmith, Scheming AIs: Will AIs fake alignment during training in order to get power?, 2023; Hubinger et al., Sleeper agents: Training deceptive LLMs that persist through safety training, 2024; Kinniment et al., Evaluating language-model agents on realistic autonomous tasks, 2023; Krakovna & Kramar, Power-seeking can be probable and predictive for trained agents, 2023; Turner & Tadepalli, Parametrically retargetable decision-makers tend to seek power, 2022; Turner et al., Optimal policies tend to seek power, 2023.
[33] Weidinger et al., Sociotechnical safety evaluations of generative AI systems.
[34] Anderljung et al., Towards publicly accountable frontier LLMs: Building an external scrutiny ecosystem under the ASPIRE framework, 2023.
[35] Bucknall & Trager, Structured access for third-party research on frontier AI models, 2023.
[36] HM Government, National Risk Register, 2023.
[37] Anderljung & Hazell, Protecting society from AI misuse: When are restrictions on capabilities warranted?, 2023.
[38] Christiano et al., Deep reinforcement learning from human preferences, 2017; Ziegler et al., Fine-tuning language models from human preferences, 2019; Lampert et al., Illustrating reinforcement learning from human feedback (RLHF), 2022; Ouyang et al., Training language models to follow instructions with human feedback, 2022.
[39] Bai et al., Constitutional AI: Harmlessness from AI feedback, 2022.
[40] DSIT, Emerging processes for frontier AI safety, 2023; Clifford, Preventing AI misuse: Current techniques, 2023.

training and education on AI risk,[41] staged release,[42] and human-in-the-loop structures.[43] To judge what constitutes adequate mitigation measures, we refer to our comments under Govern.

- **Developers of dual-use foundation models should adequately respond to incidents involving deployed models.** For particularly severe incidents, this may involve rolling back the system to a previous version or pausing operation while the incident is investigated. Developers may need to take preparatory actions, such as notifying customers about the situations under which systems will be rolled back and ensuring that there are fall-back options for e.g. critical infrastructure users.[44]

- **Developers of dual-use foundation models should track risks and mitigation measures in a risk register.** Developers of dual-use foundation models should maintain a risk register to ensure they keep track of changes in the risk landscape. The risk register should build on a risk taxonomy to ensure all important risks are covered. It should also state what mitigation measures have been taken, and who is responsible for the risk.

Guidance and benchmarks for evaluating and auditing AI capabilities

Sec. 4.1(a)(i)(C) of the E.O. 14110 directs NIST to create guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities and limitations through which AI could be used to cause harm. We recommend that:

- **Developers of dual-use foundation models should conduct model evaluations as part of their risk management activities.** For more information, we refer to our comments under Measure.

- **Developers of dual-use foundation models should develop benchmarks for dangerous model capabilities.** Model evaluations play a crucial role in responsible development and deployment of dual-use foundation models. They can inform assessments of risk, assessments of adequate safeguards, and ensure that downstream developers can use the system responsibly. The science of model evaluation is still in its early stages, both in terms of evaluation design and in terms of evaluation choice and interpretation. The mapping between evaluation results and downstream risk is still not clear. The companion resource should recommend that developers of dual-use foundation models develop benchmarks for dangerous model capabilities and support others to do so. It should encourage developers of dual-use foundation models to share best practices and the results of model evaluations unless this would be harmful.[45]

- **Dual-use foundation models should be scrutinized by external actors.** With the increasing integration of dual-use foundation models into society and the economy, decisions related to their training, deployment, and use have far-reaching implications. These decisions should not be left solely to developers. External actors need to be involved both to uncover information about the models—their risks, capabilities, and

[41] NIST, Artificial intelligence risk management framework (AI RMF 1.0), 2023.
[42] Solaiman et al., Release strategies and the social impacts of language models, 2019.
[43] Benedikt et al., Human-in-the-loop AI in government, 2020. See also Brundage et al., The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 2018; Anderljung & Hazell, Protecting society from AI misuse: When are restrictions on capabilities warranted?, 2023; DSIT, Emerging processes for frontier AI safety, 2023; Clifford, Preventing AI misuse: Current techniques, 2023; Ji, Golstein, & Lohn, Controlling large language model outputs: A primer, 2023.
[44] For a survey of measures that developers can take to respond to incidents, see O'Brien et al., Deployment corrections: An incident response framework for frontier AI models, 2023.
[45] For an overview of the current state of model evaluations, see Chen et al., Evaluating large language models trained on code, 2021; Perez et al., Discovering language model behaviors with model-written evaluations, 2022; Liang et al., Holistic evaluation of language models, 2022; Gehrmann et al., Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text, 2022.

features—but also to hold developers accountable, incentivizing them to conduct thorough evaluations and to act upon the results accordingly. Scrutiny should be applied to models themselves, but also the governance processes involved in their production and for applications built on top of them.[46] Further, achieving effective external scrutiny is no easy feat. It requires, at minimum, sufficient access to the models,[47] searching attitude on the part of the scrutinizers, proportionality to risks, independence, resources, and expertise.[48] The companion resource should include standards on how to achieve sufficient external scrutiny.

Guidelines for conducting AI red-teaming tests

Sec. 4.1(a)(ii) of the E.O. 14110 directs NIST to establish guidelines, including appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models, for conducting AI red-teaming tests. We recommend that:

- **Developers of dual-use foundation models should conduct red-teaming tests as part of their risk management activities.** For more information, we refer to our comments under Measure.
- **Developers of dual-use foundation models should advance the science of red-teaming.** Red-teaming plays an important role in uncovering issues with dual-use foundation models. It can reveal unknown model capabilities and ways in which safeguards against misuse may fail or be circumvented. Red-teaming techniques and processes can still be much improved. In particular, methods need to be developed to securely give sufficient access to external red-teamers (e.g. privacy-enhancing technologies).[49] The companion resource should recommend that developers of dual-use foundation models advance the science of red-teaming and support others to do so. It should encourage developers of dual-use foundation models to share best practices and the results of red-teaming tests unless this would be harmful.[50]

2. Reducing the Risk of Synthetic Content

Report on standards, tools, methods, and practices related to synthetic content

Sec. 4.5(a) of the E.O. 14110 directs the Secretary of Commerce to submit a report to the Director of the Office of Management and Budget (OMB) and the Assistant to the President for National Security Affairs identifying existing standards, tools, methods, and practices, along with a description of the potential development of further science-backed standards and techniques for reducing the risk of synthetic content from AI technologies.

- **Synthetic content may present significant risks.** Synthetic content will play a significant role in AI's contribution to productivity and creativity alike, from text, images, audio, and video. However, such content

[46] Mökander et al., Auditing large language models: A three-layered approach, 2023.
[47] For more details on access requirements for different kinds of safety research, see Bucknall & Trager, Structured access for third-party research on frontier AI models, 2023.
[48] For more information, see Anderljung et al., Towards publicly accountable frontier LLMs: Building an external scrutiny ecosystem under the ASPIRE framework, 2023.
[49] Bluemke et al., Exploring the relevance of data privacy-enhancing technologies for AI governance use cases, 2023; Bucknall & Trager, Structured access for third-party research on frontier AI models, 2023; Trask et al., Beyond privacy trade-offs with structured transparency, 2020.
[50] For more information on the current state of red-teaming, see Anthropic, Frontier threats red teaming for AI safety, 2023; Ganguli et al., Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022; OpenAI, GPT-4 technical report, 2023; Perez et al., Red teaming language models with language models, 2022; Touvron et al., Llama 2: Open foundation and fine-tuned chat models, 2023.

could also be used to influence elections or to conduct cyber attacks. Synthetic content from open-sourced models could present particular risks given the ease with which installed safeguards can be removed.[51]

- **Developers of generative AI should develop and distribute tools and methods to identify synthetic content.** One way to mitigate the risks from synthetic content is to ensure that it can be identified as such. That way, social media platforms could tag AI-generated content, giving citizens the ability to make more informed choices. It can also aid in ensuring that citizens are made aware if they are interacting with an AI-based chatbot rather than a human. Detectors of AI-generated content appear to be low accuracy,[52] and will likely need to be aided by interventions by developers. Some such efforts are already underway.[53]

  Developers of generative AI can introduce content provenance tags and watermarks into their content. Content provenance tags encodes details about the content's provenance using cryptography into its metadata.[54] This technique is helpful for images, audio, and video as they are distributed in files that already contain metadata. However, it is of limited use for text, which typically is not accompanied by metadata. Furthermore, content provenance tags may not be robust to adversarial attempts, e.g. they can often be removed by taking a screenshot of the relevant picture. As such, content provenance tags are more appropriate for cases where actors are incentivized to prove rather than hide the authenticity of content, such as a government body seeking or a journalist wishing to prove the authenticity of a picture.

  Given the limitations of content provenance techniques, they should be supplemented with watermarks. Techniques for such watermarks involve introducing humanly imperceptible biased noise in the content which a classifier could identify, and are reasonably mature in audio, video, and images. Such watermarks are yet to see widespread adoption in the market. We hypothesize that this is because generative AI companies expect that using watermarks would put them at a competitive disadvantage and because of uncertainties regarding their effectiveness. While introducing watermarks for AI-generated text is significantly more challenging and appears less robust to attempts at removal, they deserve significant study. Despite some limitations,[55] we expect such watermarks to be sufficiently effective to be worthwhile. As such they deserve significant study, investment, and development.[56]

- **Developers of generative AI should develop and distribute other tools to address the risks from synthetic content.** Tools beyond identifiers of whether some piece of content is AI-generated or not only present one of several interventions to address risks from synthetic content. Developers of generative AI are in a particularly good position to develop such tools, given their technical capabilities and their greater insight into model capabilities and use. They may also have a unique responsibility to do so, as such tools would address externalities imposed by technologies they've developed and marketed.

[51] Seger et al., Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives, 2023.
[52] Though there is recent non-peer reviewed or replicated research which is claimed to yield impressive results, Hans et al., Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text, 2024.
[53] For example, OpenAI announced they would introduce content provenance techniques into their generated images ahead of 2024 elections, see OpenAI, How OpenAI is approaching 2024 worldwide elections, 2024. Similarly, Google DeepMind announced that their music generation model Lyria would include inaudible watermarks, see Google DeepMind, Transforming the future of music creation, 2023.
[54] The Coalition for Content Provenance and Authenticity (C2PA) is developing standards for such metadata, see C2PA, Introducing official content credentials icon, 2023.
[55] Zhang et al., Watermarks in the sand: Impossibility of strong watermarking for generative models, 2023.
[56] For promising directions, see Kirchenbauer et al., A watermark for large language models, 2023; Zhao et al., Provable robust watermarking for AI-generated text, 2023..

3. Advance Responsible Global Technical Standards for AI Development

Plan for global engagement on promoting and developing AI consensus standards, cooperation, and coordination

Sec. 11(b) of the E.O. 14110 directs the Secretary of Commerce to establish a plan for global engagement on promoting and developing AI consensus standards, cooperation, and coordination, ensuring that such efforts are guided by principles set out in the NIST AI Risk Management Framework (AI RMF)[57] and the U.S. Government National Standards Strategy for Critical and Emerging Technology.[58] Against this background, we recommend to:

- **Engage with the increasing number of AI Safety Institutes globally.** Besides the U.S., the UK is the only other country that has already set up an AI Safety Institute.[59] But we expect other countries to do the same over the coming months (e.g. Japan[60] and Canada[61]). These institutes will likely inform standard-setting and policy-making with regards to the most capable AI models globally. NIST should therefore engage with these institutes, presumably via the U.S. AI Safety Institute.

- **Coordinate with European standard-setting processes.** The EU is about to implement the world's first comprehensive AI regulation. Given that most developers of dual-use foundation models will want to offer their products in the EU—due to the size of its market—it may prove cheaper to offer EU-compliant products in the U.S. rather than maintaining two separate product lines. We therefore expect a Brussels Effect for certain requirements in the AI Act, especially with regards to dual-use foundation models that require significant capital investment.[62] Against this background, it seems particularly important that NIST engages with the European Standardization Organizations (ESO)—especially CEN/CENELEC Joint Technical Committee 21 on AI—as well as the relevant parts of the European Commission that will refine and clarify requirements imposed by the AI Act. This will include the soon-to-be established AI Office that is tasked with developing and enforcing the EU's rules on "general-purpose AI (GPAI)"—the term used in the AI Act to refer to foundation models.

- **Participate in international standard-setting processes.** International standard-setting organizations are currently developing standards on AI, including generative AI and dual-use foundation models. Import standard-setting organizations include ISO, IEC, and IEEE. NIST should make sure to participate in international standard-setting processes. If international standards diverge substantially from domestic ones, then this might impose extra burdens on American developers.

[57] NIST, Artificial intelligence risk management framework (AI RMF 1.0), 2023.
[58] The White House, U.S. government national standards strategy for critical and emerging technology, 2023.
[59] DSIT, Introducing the AI Safety Institute, 2023.
[60] The Yomiuri Shimbun, Japan government to establish AI Safety Institute in January, 2023.
[61] Hemmadi, Bengio backs creation of Canadian AI safety institute, will deliver landmark report in six months, 2023.
[62] For more information, see Siegmann & Anderljung, The Brussels Effect and artificial intelligence: How EU regulation will impact the global AI market, 2022.

Further resources

GovAI researchers have published several pieces relevant to this RFI:

AI risk management:

- Schuett, Risk management in the Artificial Intelligence Act, 2023
- Koessler & Schuett, Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries, 2023
- Schuett, Three lines of defense against risks from AI, 2023
- Schuett, AGI labs need an internal audit function, 2023
- Schuett et al., How to design an AI ethics board, 2024
- Mulani & Whittlestone, Proposing a foundation model information-sharing regime for the UK, 2023
- Brundage et al., The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 2018
- Anderljung & Hazell, Protecting society from AI misuse: When are restrictions on capabilities warranted?, 2023
- Shevlane, Structured access: An emerging paradigm for safe AI deployment, 2022
- Seger et al., Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives, 2023
- Schuett et al., Towards best practices in AGI safety and governance: A survey of expert opinion, 2023
- Alaga & Schuett, Coordinated pausing: An evaluation-based coordination scheme for frontier AI developers, 2023
- Schuett & Anderljung, Comments on the initial draft of the NIST AI risk management framework, 2022

Evaluations, auditing, and red-teaming:

- Shevlane et al., Model evaluation for extreme risks, 2023
- Anderljung et al., Towards publicly accountable frontier LLMs: Building an external scrutiny ecosystem under the ASPIRE framework, 2023
- Mökander et al., Auditing large language models: A three-layered approach, 2023
- Bluemke et al., Exploring the relevance of data privacy-enhancing technologies for AI governance use cases, 2023
- Trask et al., Beyond privacy trade-offs with structured transparency, 2020
- Bucknall & Trager, Structured access for third-party research on frontier AI models, 2023
- Rando et al., Red-teaming the stable diffusion safety filter, 2022

AI regulation:

- Anderljung et al., Frontier AI regulation: Managing emerging risks to public safety, 2023
- Schuett, Defining the scope of AI regulations, 2023
- Schuett, Risk management in the Artificial Intelligence Act, 2023
- Smith et al., Response to the NTIA AI accountability policy request for comment, 2023