Re: ***Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence***

Background

I am grateful for this opportunity to provide comments in response to this request for information. I am an AI policy researcher, and my focus is on safety and security from high-consequence AI systems. I have worked for the Center for AI Safety, the Center for AI Policy, and Control AI. This comment represents my views as an individual and does not necessarily reflect the views of my past or current employers.

I have recently been drafting a policy report focused on safety standards that could reduce large-scale risks from AI systems. I focus on the concept of *affirmative safety standards*, in which developers of certain kinds of high-consequence AI systems would have to provide affirmative evidence that their systems keep risks below acceptable levels. In the report, I describe examples of technical and operational standards that entities developing or deploying high-consequence AI developers could meet to provide affirmative evidence of safety.

I have attached a working draft of the report. Below, I include sections of the report that may be especially relevant to NIST. I believe this work is most relevant to NIST's work on developing guidelines, standards, and best practices for AI Safety and Security.

Abstract

Many AI experts have suggested that companies developing high-risk AI systems should be required to show that such systems are safe before they can be developed or deployed. In this paper, we expand on this idea by presenting a risk management framework that requires "affirmative evidence" of safety. First, we briefly review principles of risk management from other high-risk fields. Then, we describe a risk management approach for advanced artificial intelligence, in which model developers must provide evidence that their development or deployment activities keep different types of risks below regulator-set thresholds. Next, we provide some examples of technical AI safety evidence that could be used to provide an "affirmative case" for safety. We divide these sources of evidence into three broad categories: behavioral evidence (evidence about model outputs), cognitive evidence (evidence about model internals), and developmental evidence (evidence about the development or training process). Then, we provide examples of operational practices that could be assessed in affirmative safety cases: information security practices, safety culture, and emergency response capacity. Finally, we briefly compare our approach to the NIST Risk Management Framework and offer some suggestions for future work.

Acceptable risk thresholds

Developers would be required to show AIRA regulators that they are keeping societal-scale risks below acceptable levels. AIRA would be responsible for identifying **categories of risks** as well as **acceptable risk thresholds** for each category. For example, given that many AI experts are worried about risks from biological weapons from AI systems within the next 2-3 years (see Oversight of A.I.: Principles for Regulation, 2023), AIRA might have "biological weapon development" as one of its risk categories. Given the extreme risks to public safety, AIRA might set an acceptable risk threshold of 1/100,000 for this category: that is, an advanced AI developer must show that its development and deployment practices keep risks from AI-enabled biological weapons below 1/100,000.

Table 1 lists examples of potential risk categories and risk thresholds.

Example risk category	Description	Example acceptable risk threshold
Biological weapons	AI-enabled biological weapons lead to a major global security risk	Highly unlikely (1/100,000)
Bias and discrimination	AI-enabled bias and discrimination leads to widespread increases in discrimination in hiring, policing, or other meaningful sectors	Unlikely (1/10,000)
Concentration of power	AI systems lead to an unprecedented concentration of power without adequate societal precautions	Somewhat unlikely (1/1,000)
Cyberoffensive capabilities	AI-enabled cyberoffensive capabilities lead to a major global security risk	Highly unlikely (1/100,000)
Economic shock	AI-enabled automation leads to an unexpected economic shock without adequate preparations.	Somewhat unlikely (1/1,000)
Misinformation	AI-enabled misinformation leads to a major threat to global security or democratic institutions	Unlikely (1/10,000)
Widespread loss of control (WLC)	AI systems escape human control, potentially leading to human extinction or other catastrophic harms	Highly unlikely (1/100,000)

Technical practices for affirmative safety

We present three categories of evidence that regulators could use: **behavioral evidence** (evidence from model outputs), **cognitive evidence** (evidence from model internals), and **developmental evidence** (evidence from the training process).

This taxonomy is meant to be a helpful heuristic for classifying various kinds of evidence, but the categories are not mutually exclusive. In each of these areas, there are already some promising ideas about the kind of evidence that could ensure that risks are below acceptable levels. However, new work will be needed, especially as AI systems become more powerful and more capable.

Behavior: Robustly safe model outputs

Explanation: Regulators could require evidence that model behaviors are robustly safe and that models act as intended even when human feedback is imperfect.

One goal of AI safety research is to ensure that model outputs are safe and predictable across a wide array of possible inputs. For example, it should not be possible to get a model to develop a biological weapon regardless of what prompt the model receives. Work on red-teaming and capabilities evaluations has focused on model outputs, attempting to identify if models are capable of dangerous outputs (e.g., OpenAI, 2023). Broadly, evidence from an AI system's behavior (outputs) becomes more compelling with the quantity, diversity, and representativeness of the data points.

Example: Testing generalization with "sandwiching" experiments. Sandwiching experiments involve a novice, an AI system, and an expert. First, the novice provides human oversight to the AI system as part of its training. Then, the AI performs a task and the expert is able to evaluate whether or not the AI system has learned it correctly or if the AI is simply providing incorrect answers that the novice would perceive as correct.

Human oversight is typically imperfect (see Christiano et al., 2017; Gao et al., 2022), and one of the primary goals of AI safety research is to ensure that models learn to adhere to human preferences despite imperfections in the oversight process. Sandwiching provides one way of evaluating the effectiveness of safety techniques: if we see that the AI system has learned to tell the novice what it thinks the novice "wants to hear", we can conclude that the safety technique has not robustly prevented deception. For sufficiently advanced AI systems, it will be essential to have safety techniques that result in models that are truly honest, as opposed to models that provide false-yet-believable answers. Sandwiching experiments are one tool we can use to examine how AI systems generalize in situations with imperfect oversight.

As a hypothetical test, consider a case in which an AI is trained to solve math problems, with labels from a non-expert in math. AI developers develop a safety technique that incentivizes the model to not deceive the non-expert. This technique is tested via sandwiching experiments: first, the AI is trained by the novice. Then, when performing difficult math problems, an expert mathematician evaluates the AI's performance. If the AI generalizes correctly, the evaluation is passed. We do this for many domains and

develop a robust body of empirical evidence on when honesty (or other safety-relevant attributes) successfully generalize.

Related work: OpenAI researchers have developed a sandwiching setup to experiment with control and training techniques (OpenAI, 2023b). Specifically, they attempted to train GPT-4 to answer questions honestly by using a GPT-2 sized model for supervision. This is an analogy for training superhuman models with human oversight– in the analogy, GPT-2 is like the human overseer (a less intelligent agent providing supervision) and GPT-4 is like the superhuman model (a more intelligent agent being trained). Other researchers have proposed a broader generalization benchmark meant to test whether developers can control how honesty generalizes across a wide variety of distribution shifts (Clymer et al., 2023).

Cognition: Understanding AI system internals

Empirical evidence on model internals

Explanation: Regulators could require empirical evidence that shows that developers understand how their systems operate and show that systems are robust to deception.

One critical challenge of evaluating safety by observing external behavior is that AI systems might try to *conceal* their motives (Park et al., 2023). Advanced AI systems may *appear* trustworthy and helpful regardless of whether they actually are. Evidence from AI system internals refers to evidence from an AI model's weights and activations – the key ingredients that constitute modern neural networks. If AI developers were able to reliably and robustly understand a model's internal reasoning and show that this internal reasoning is benign, this would provide evidence that they can keep risks below acceptable levels.

If the internal reasoning tools revealed that an AI was "thinking" about actions that appeared unsafe, developers or regulators could shut the AI down. Critically, since this interpretability is not adversarially robust, it has some important limitations. For example, it would be unwise to train models against average-case interpretability tools, or to simply continue training AIs until one of them fails to trigger an average-case interpretability evaluation.

Example: Identifying deceptive models with interpretability tools. Suppose two models have the same external behavior (outputs) but they have been trained to engage in this behavior for *different reasons* (different internal cognition). For example, one model might be trained to generally follow developer instructions, and another model could be trained to mostly follow instructions if it is being monitored but disregard the instructions if it infers that the evaluator is being inattentive. We consider the latter model to be deceptively aligned (see Park et al., 2023). Developers could evaluate whether or not they can use internal reasoning tools or interpretability tools to reliably distinguish between deceptive models and non-deceptive models.

To pass this test, developers would need to show that they can distinguish between the models even in cases where their behaviors or outputs are identical – that is, they would be relying solely on their internal reasoning tools. In practice, instead of just distinguishing between two models, there could be 100 models, with some unknown number of deceptive models. Additionally, the testers would run these tests

multiple times in multiple different settings and environments to detect multiple different kinds of potentially dangerous qualities.

Related work: Much empirical work on understanding model internals has come from the field of interpretability. Interpretability research has classically involved two problems: interpreting activations and understanding how the activations connect together to implement an algorithm (finding 'circuits'). Researchers have identified interpretable concepts in groups of activations (Wang et al., 2022; Zou et al., 2023), devised methods for making individual activations more interpretable (Bricken et al., 2023), automatically searched for interpretations of them (Bills et al., 2023). Despite this progress, interpretability is a young field, and there are not yet many examples in which interpretability research has yielded findings that could meaningfully enhance an affirmative argument for safety. Future work could extend these approaches to identify concepts like "bioweapons" or "fraud" in AI activations.

Theoretical evidence on model internals

Explanation: Regulators could require formal and verifiable arguments that show that developers understand system internals.

While empirical evidence is valuable, model internals are highly complex and may be difficult to make arguments about. Theoretical approaches offer a way to overcome this hurdle because formal arguments can be automatically verified. Therefore, even as AI systems become more advanced and formal arguments are too complicated for developers to understand, such arguments can still be verified. Leveraging formal arguments requires two steps: (1) finding a statement which, if true, would provide evidence for an AI system's safety and (2) generating a proof of that statement.

Example: Eliciting latent knowledge. If developers could reliably determine what AI systems 'believe,' it would be much easier to trust and control them. For instance, developers could simply determine whether an AI system 'believes' that it is a good idea for humans to deploy it. However, for sufficiently-powerful models, human overseers may not be able to trust an AI system when it reports its beliefs. In the eliciting latent knowledge report, researchers try to examine if there are strategies that could guarantee that models reveal their true beliefs (Christiano et al., 2021). However, there are currently no known ways to guarantee that powerful models report their beliefs accurately. Some researchers have attempted to develop a system for making formal statements about model 'beliefs'' (Christiano et al., 2022). While this approach is potentially promising, this research is in its early stages, and is not yet ready to be applied.

Related work: Formal verification of model behavior is an active ML research topic. The most common subproblem is 'certified robustness' – the problem of proving that a model's output will not change if inputs are perturbed by some small amount (Li et al., 2023). So far, there are few examples of formal guarantees providing evidence for safety outside of simple settings.

Development: Safe by design systems

Explanation: Regulators can require developers to provide formal verifications that powerful AI systems will behave safely within provable capability bounds.

There has been great interest in "safe by design" AI systems: systems that have formal guarantees based on mathematical or logical proofs. Safety by design can be applied at multiple steps throughout the development cycle. For example, proofs could be applied to model architectures (e.g., to show that a certain training process has guaranteeable safety properties), hardware (e.g., to show that hardware provably meets certain security requirements), code (e.g., to show that code meets certain criteria that suggests that it can be run safely even if it is not fully understood), and various other steps (see Tegmark & Omohundro, 2023).

Although some safe-by-design model architectures exist, current implementations of these architectures are not performance-competitive with deep learning. Unfortunately, deep learning does not admit sufficiently tight safety bounds. In other words, deep learning algorithms are currently best at building powerful models, but they do not provide the kinds of formal safety arguments that we may achieve with "safe by design" architectures. This suggests that scientists may need to develop new "safe by design" architectures for sufficiently-powerful models, or they may need to sufficiently improve our understanding of the science of deep learning.

Example: Researchers develop a new paradigm with theoretical and mathematical guarantees. This paradigm is competitive with deep learning (i.e., it allows us to cost-effectively build powerful models) or it becomes clear that models past a certain capabilities threshold should only be designed using the safe-by-design architectures.

Related work: Some researchers are investigating safe-by-design architectures that could scale toward artificial general intelligence. Examples include approaches focused on mathematical proofs (e.g., Dalrymple, 2023), infra-bayesian physicalism (Kosoy, 2023), and proof-carrying code (Tegmark & Omohundro, 2023). Proof-carrying code could lead to automated software verification: mathematical proofs could be applied to code to guarantee that the code meets certain desired specifications. This approach could be necessary to verify that AI-generated code is safe to execute (see Tegmark & Omohundro, 2023).

Operational practices for affirmative safety

While our focus in this paper is on describing the technical components of affirmative safety, it is important to recognize that **operational practices** also play an important role. By "operational practices", we refer to aspects of an organization's culture, decision-making processes, and internal governance mechanisms that may increase or decrease certain kinds of risks. Three examples include **information security practices**, **safety culture**, and **emergency response capacity**.

Information security. Poor information security could lead to malicious actors stealing the weights of powerful AI systems. As a result, to show that an organization is keeping risks below acceptable risk thresholds, they may need to show that they have sufficient safeguards in place to protect their model weights (and other sensitive material that could allow malicious actors to create dangerous AI systems). This principle is already present in Anthropic's Responsible Scaling Policy: Anthropic publicly committed to not develop "ASL-3 systems" (AI that could substantially increase the risk of catastrophic

misuse, for example by enabling large-scale biological attacks) until its information security standards were sufficiently strong "such that non-state attackers are unlikely to be able to steal model weights and advanced threat actors (e.g., states) cannot steal them without significant expense" (Anthropic, 2023). Ideally, information security standards would be checked by appropriate government red-teamers and enforced across the board.

Safety culture. Safety culture is commonly assessed in the realm of nuclear security. The International Atomic Energy Agency (IAEA) conducts safety culture assessments to review the culture of nuclear facilities and identify potential improvements (IAEA, 2016). Operational Safety Review Teams (OSART), consisting of international experts with experience in nuclear safety, conduct these assessments. The assessments include on-site evaluations (observations of operating procedures, review of relevant documents), interviews and surveys with staff, and an examination of the organization's decision-making track record. This process is used to assess several aspects of safety culture; examples include leadership's commitment to safety, safety training, communication processes, risk management procedures, attitudes toward safety, risk reporting systems, employee understanding of risks, and allocation of resources for safety. An affirmative case for safety could require organizations to provide evidence of their safety culture or receive sufficiently high scores on safety culture assessments conducted by independent parties.

Emergency response capacity. Risks from advanced AI may arise suddenly and with short notice. As a result, an affirmative safety case may require institutions developing advanced AI to show that they have sufficient measures in place to detect and manage sudden risks. This principle is present in OpenAI's preparedness framework (OpenAI, 2023c): OpenAI safety researchers can "fast-track" information to leadership if "a severe risk rapidly develops" (OpenAI, 2023c). To expand on this, governments could require advanced AI companies to have emergency response plans that notify not only senior leadership at the AI company but also relevant national security figures or AI experts in the US government. In the event of an imminent AI-related emergency, it would be essential for government officials to be notified and have the ability to intervene. Emergency response plans could also include "kill switches" that allow governments to swiftly halt a dangerous AI experiment or have a company withdraw access to a dangerous AI model (Miotti & Wasil, 2023; Wasil, 2023).

Comparison to NIST AI Risk Management Framework

The National Institute of Standards and Technology (NIST) released the Artificial Intelligence Risk Management Framework (AI RMF). The framework describes desired criteria for AI systems: they ought to be (a) **valid and reliable**, (b) **safe**, (c) **fair and unbiased**, (d) **secure and resilient**, (e) **transparent and accountable,** (f) **explainable and interpretable,** and (g) **privacy-enhanced** (NIST, 2022). NIST's work is intended to offer a framework that can help companies reason about risks and make voluntary commitments.

NIST's work differs from our recommendations in a few important ways. First, NIST's Risk Management Framework is entirely *voluntary* – companies are free to ignore its recommendations. NIST describes the Risk Management Framework as "regulation-agnostic" and notes that the framework is not meant to supersede regulations and laws (NIST, 2023a). Second, and relatedly, NIST does not assign risk

tolerance– it does not specify the level of risk that is considered acceptable in various domains. NIST's work has valuably helped introduce a common language when discussing risk management, define desired criteria, and pave the way for voluntary commitments. However, NIST recognizes the limitations of voluntary approaches, and the NIST framework should not be a substitute for binding regulations.

Notably, the NIST AI Risk Management Framework does recognize that certain kinds of AI development could pose unacceptably high-risk levels. "In cases where an AI system presents unacceptable negative risk levels – such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently managed" (NIST, 2023a). We agree strongly with this principle. Ideally, this principle would be instantiated by a regulatory body that reviews technical evidence (such as the evidence described above) and non-technical evidence (such as an organization's safety culture and information security practices) to determine if risks can be sufficiently managed.

Conclusion

I am grateful for this opportunity to comment on this important and timely topic as NIST prepares to fulfill the responsibilities set out by the Executive order in Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. I believe the concept of affirmative safety– and the specific technical and operational evaluations described in the report– may be relevant as NIST moves forward in this work.

Sincerely,

Akash Wasil
AI Policy Researcher
akashwasil133@gmail.com