Workshop Report

# Repurposing the Wheel:

## Lessons for AI Standards

**Authors**
Mina Narayanan
Alexandra Seymour
Heather Frase
Karson Elmgren

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

November 2023

## Executive Summary

Standards are crucial in ensuring smooth market function, interoperability, and consumer safety, as well as aiding in the development of regulation for new technology. However, establishing standards for rapidly evolving artificial intelligence (AI) technologies is complex, due to challenges including the absence of universal definitions surrounding AI and the explosion of potential AI use cases. The family of related AI technologies presents societal risks that require various levels of oversight and a nuanced approach to standard-setting.

A series of workshop sessions co-organized by the Center for Security and Emerging Technology and the Center for a New American Security in the fall of 2022 examined case studies of previous standards development across several industries to draw lessons for AI. These discussions highlighted the challenges of developing robust, effective standards as well as best practices that have enhanced standards development and enforcement processes in the domains we studied. The workshop and many of its recommendations were completed before the October 30, 2023 release of the Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The recommendations are consistent with the Executive Order and frequently provide details and specificity for implementing it.

Our key findings are:

- **Finding 1: AI risk assessment and mitigation should include examining how interdependencies affect systemic risk.**
  - *Recommendation 1:* Critical infrastructure owners and operators should track the interdependencies of their AI systems.
  - *Recommendation 2:* The Office of Management and Budget (OMB) and the Office of Science and Technology Policy (OSTP) Directors' forthcoming guidance on minimum risk management practices for AI should require that agencies identify the risks that could emerge from interdependencies between their AI systems and other entities.
- **Finding 2: *Guidance on testing and re-approval of AI systems should be calibrated to risk and account for changes to AI systems over time.***
  - *Recommendation 3:* The U.S. Department of Defense should create thresholds or triggers for different levels of rigor and oversight for testing military AI systems.

    - *Recommendation 4:* U.S. government agencies should establish processes for the reassessment and re-testing of systems as they change over time and share these processes with each other.
- **Finding 3: Compliance assistance can help small- and medium-sized businesses prepare for and implement AI regulation.**
    - *Recommendation 5:* Congress should create a pilot AI Compliance Assistance Office within the U.S. Department of Commerce, which should later expand to other government agencies.
- **Finding 4: Third-party organizations can remove barriers to standards development, implementation, compliance, and tracking.**
    - *Recommendation 6:* OMB should direct a study by an independent body to inform the designation of third-party accreditation bodies that ensure certifiers evaluate the implementation of AI standards in a consistent manner.
    - *Recommendation 7:* Professional organizations should establish AI standards access funds, whistleblower protection programs, and reporting programs to gather anonymized information on AI risks from industry participants.
- **Finding 5: Non-regulatory governance is one mechanism that can support the safe development and use of AI systems.**
    - *Recommendation 8:* The United States should commence discussions in the G7 about creating the equivalent of a Financial Action Task Force for AI.
    - *Recommendation 9:* NIST should create an online portal to ensure technical developments relevant to standards are captured and publicized.
- **Finding 6: Coordination and regular efficacy checks of standards can ensure that standards development is efficient and effective.**
    - *Recommendation 10:* Standard-setting bodies should host biannual summits to coordinate on standards interoperability and efficacy.
    - *Recommendation 11:* NIST should support the development of testbeds to monitor AI standards for effectiveness.

## Table of Contents

## Introduction

Standards are "the common and repeated use of rules, conditions, guidelines, or characteristics for products or related processes, practices, and production methods."[1] By providing a common language for products or approaches, standards enable interoperability and trust in technologies, helping markets to function smoothly. Standards support consistent performance measurement and evaluation, promote interoperability between components from different firms, and protect consumers by ensuring safety, durability, and market equity.[2] By assisting with finding consensus on best practices among technical experts, standards can also guide regulators and governments in developing regulation for new technologies.[3]

Standards compliance is often achieved using conformity assessments, which are defined by the International Organization for Standardization (ISO) and the International Electrotechnical Commission, two international standards bodies, as a "demonstration that specified requirements are fulfilled" by way of testing, inspection, auditing, certification, and accreditation.[4] These tasks may vary depending on the form of the standard. Standards can be qualitative or quantitative, regulatory or voluntary, industry- or government-led, national or international, or a combination of any of these. Some standards are documentary, meaning that they provide an approved way to carry out a technical process.[5] Others are measurement based, meaning they embody a quantity and define a unit, such as a kilogram or a meter.[6] Consequently, standards are a key component of promoting good governance practices domestically and internationally, which is essential for driving innovation responsibly.

*Standards are a key component of promoting good governance practices domestically and internationally, which is essential for driving innovation responsibly.*

While governments recognize the importance of having enforceable standards, artificial intelligence (AI) standards are still in a nascent stage. AI systems promise to positively impact people's lives by performing or augmenting certain functions that humans struggle with, such as detecting patterns in or extracting high-value information from large amounts of data. However, these systems can also create harm in a number of ways – for instance, by propagating misinformation or

reinforcing biases inferred from training data. Although standards can enable consistency and conformance with practices that reduce risks and promote beneficial outcomes of AI systems, they have been slow to emerge for two main reasons:

1. AI lacks a universally agreed upon definition, making it difficult to develop widely-accepted measures and metrics for AI.[7] This report recognizes the multiple existing definitions of AI and offers recommendations that would align with many of these without preferring one in particular.
2. AI technologies are rapidly evolving and application-specific, which means that standards tied to a specific AI system architecture may not be useful for long or be applicable to a wide range of systems. The design of standards would also need to account for the risk variance of environments AI systems are integrated in, ranging from low-stakes use cases like online shopping to high-stakes applications such as weapons systems.

In short, the AI landscape is complex. It brings various societal risks that require different levels of oversight, depending on the use case. Moreover, AI is not a single technology, but a family of related technologies, meaning it must be treated with nuance in standard-setting. Fortunately, developing standards is not a new challenge; ISO alone has over 24,000 international standards.[8] As a result, stakeholders involved in the AI standards development process can examine other domains with well-defined standards to learn from their history.

### *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*

On October 30, 2023, President Biden signed an Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The Executive Order focuses on "governing the development and use of AI safely and responsibly" and "advancing a coordinated, Federal Government-wide approach to doing so." This lengthy document contains over 100 provisions* that seek to ensure the safety and security of AI systems; enable the United States to lead in promoting responsible AI innovation, competition, and collaboration; protect U.S. worker rights; promote equity and civil rights in AI policies, protect consumers; support data protection and privacy; manage the risks of AI for service delivery in the U.S. government; and support U.S.

---

* CSET has a publicly available spreadsheet that tracks the provisions in the executive order. https://cset.georgetown.edu/article/eo-14410-on-safe-secure-and-trustworthy-ai-trackers

government leadership in AI via multilateral engagements. Of the content in the Executive Order, sections 4 ("Ensuring the Safety and Security of AI Technology") and 10 ("Advancing Federal Government Use of AI") are most relevant to this document and are often referenced in our recommendations.

### *Office of Management and Budget (OMB) Draft Policy on the Use of AI in the Federal Government*

The Office of Management and Budget (OMB) released a draft guidance memorandum on November 1 to support the commitments made in the October 30 Executive Order. The draft policy strengthens AI governance at the federal level through three main initiatives.

- The first initiative aims to develop AI governance structures by establishing federal agency leadership positions and coordinating bodies for AI.
- The second initiative seeks to advance responsible AI innovation through efforts such as improving agency enterprise infrastructure and workforce capacity.
- The third initiative focuses on managing AI risks by mandating specific safeguards for rights- and safety-impacting uses of AI, which are defined in the policy, and providing recommendations for managing risk in federal procurement and contracts.

This third initiative, focusing on AI risks, contained in Section 5 of the OMB draft policy is the most relevant to our recommendations.

### *Technology Policy Lab Working Group Sessions*

To extract lessons from the rich history of standard-setting in other sectors, the Center for Security and Emerging Technology (CSET) and the Center for a New American Security (CNAS) – both think tanks that conduct research and analysis at the intersection of technology, policy, and national security – led a Technology Policy Lab Working Group to study examples where standards are well-established to identify takeaways that are relevant to AI. Over the course of five sessions between July 2022 and October 2022, invited speakers presented case studies to a selected group of stakeholders that helped illustrate the process of standards development in various sectors. From these discussions, it became clear that a roadmap for AI standards should be informed by successes and failures in other industries.

## Methodology

The Technology Policy Lab Working Group consisted of stakeholders from universities, companies, nonprofits, and U.S. and international government organizations. Participants had expertise in areas such as risk management, human-machine teaming, testing and evaluation, defense, safety, standards, policy, and security. Most had experience applying these areas to AI.* The working group analyzed five case studies of national and international standard-setting: banking institutions in the financial sector, the Occupational Safety and Health Administration's (OSHA) safety standards and enforcement mechanisms, standardization developments in the cybersecurity sector, sustainability in the building sector, and medical device approvals by the Food and Drug Administration (FDA). The next section summarizes each case study.

### *Working Group Goals*

The working group sought to achieve three overarching goals:

1. Identify insights, cautionary tales, successes, and actionable steps from standards development across the case studies to apply them to AI standards;
2. Better understand possible approaches for the development of AI standards;
3. Generate ideas for future AI standards development.

To achieve these goals, we assessed each case study's standards development process to estimate the time and effort required to complete standardization steps, determine key stakeholders needed at each step, and identify challenges and successes when establishing and implementing standards. In doing so, we hoped to learn how standards are operationalized and demonstrated systematically and repeatedly.

During the final session of the workshop series, participants synthesized key themes from the previous sessions and brainstormed areas of opportunity within AI standard-setting. From these themes, we drew several takeaways for AI standards. The Key Takeaways section discusses six overarching takeaways from our discussions and provides 11 recommendations to operationalize lessons learned as U.S. policymakers tackle AI standards development.

* For a full list of participants, please see the Acknowledgements section below.

## Case Studies

### *Banks*

The first case study of the workshop covered global standard-setting for banking institutions. The case study focused on several financial organizations that play a central role in standard-setting. The Bank for International Settlement (BIS) and the Financial Stability Board (FSB) are two institutions that enable coordination among financial institutions and define standards for promoting financial stability. BIS uses preestablished criteria in its standards to define a bank as critical to the stability of the financial system. FSB takes a more flexible approach when it comes to standard-setting. For instance, FSB allows *banks of systemic importance* to choose the proportion of instruments and liabilities that act as a buffer against bankruptcy – as long as the proportion falls within a predefined range.* Workshop participants raised several mechanisms that enhance the ability of organizations to manage risk and compel organizations to behave responsibly. The discussion following the case study highlighted how the Financial Action Task Force (FATF) uses public lists to incentivize countries to counter illicit financial activities.

### *Occupational Safety and Health Administration (OSHA)*

OSHA, part of the U.S. Department of Labor, was created by Congress in 1970 to set and enforce worker safety standards.[9] This case study featured speakers from government and industry who explained how OSHA standards are enforced and detailed the process for how new standards are proposed and approved. OSHA enforcement occurs at both the federal and state levels, with 22 states maintaining their own OSHA-approved state plans.[10] When state-level OSHAs are found to violate federal-level requirements, federal OSHA can step in and take over. The speakers highlighted how OSHA offers compliance assistance services that pause enforcement and help organizations implement OSHA requirements. Importantly, the OSHA case study provided an in-depth look at the rulemaking process, which includes seven stages and, as presenters noted, can take 10 to 12 years.[11] Participants spent most of the session critically examining each of the seven stages to understand the stakeholders, level of public participation, and length of processes involved. One noteworthy discussion point was how industry associations can submit anonymous

* Banks of systemic importance are banks whose failure could pose a threat to the international financial system.

consensus opinions for the public docket, which encourages information sharing while addressing liability concerns. The presenters noted how OSHA standards are not meant to be novel, emphasizing the critical role consensus bodies play in standards creation.

***Cybersecurity***

The cybersecurity case study was selected to glean lessons from a sector that shares many of the same challenges as AI, but is more mature in terms of U.S. government regulation, standards, workforce professionalization, and available guidance.[*] The presenter outlined documents and processes relevant to the cybersecurity standards landscape, including the U.S. National Institute of Standards and Technology's (NIST) Cybersecurity Framework, the Federal Risk and Authorization Management Program (FedRAMP), and Executive Order 14028, Improving the Nation's Cybersecurity. An important lesson from this presentation was the need to update standards incrementally because systems must be able to appropriately adapt as threats change. Participants discussed how building trust in cybersecurity systems must be an important focus as standards are updated. The presenter identified supply chain vendor trustworthiness for both software and hardware as an area of opportunity for progress. Finally, participants discussed the criticality of workforce development for the cybersecurity ecosystem, specifically examining documents such as NIST's Workforce Framework for Cybersecurity (the National Initiative for Cybersecurity Education Framework) and the MITRE ATT&CK® framework.

***Sustainability***

The sustainability case study covered standards for resource usage and environmental sustainability in the buildings sector. The discussion focused on the Leadership in Energy and Environmental Design (LEED) building standard, which requires a building to achieve certain metrics that are designed to promote ecological sustainability, from water usage to carbon emissions released over the lifecycle of building materials to provision of bicycle parking. Participants discussed the reasons for the relative success of the LEED standard, which has been widely adopted in North America and Europe, and the limitations of LEED. The presenter noted that LEED is relatively ill-suited for developing economies, ignores operational factors over the full building lifecycle, and

[*] Shared challenges include an array of dynamic risks that can extend beyond the enterprise, rapidly evolving technological methods and tactics, lack of workforce talent, adversarial contexts, and supply chain issues.

in practice can struggle to account for energy consumed by processes associated with the construction of a building. In addition to the LEED standard, participants discussed how the sustainability standards ecosystem has developed over time.

***Medical Devices***

The medical devices case study focused on established FDA processes for approval of medical devices. The case study covered the history of medical device approvals, which includes cautionary tales about devices that were inadequately tested and caused significant harm, such as the Dalkon Shield intrauterine device. Presenters described FDA's Class I, II, and III structure for evaluating medical devices based on risk of harm to patients, and different levels of uncertainty based on the presence or absence of historical data on safety and efficacy. Participants discussed special approval processes for novel and minimal-risk devices, including some of the problems with current approval processes.

## Key Takeaways

***Takeaway 1***

***AI risk assessment and mitigation should include examining how interdependencies affect systemic risk.***

Risk is defined as the adverse impacts that would result from an event combined with the likelihood of that event occurring. The ways that risk can arise are not always apparent, but should be taken into account in standards development.[12] Risks that emerge at scale, and that may result from interdependencies between entities, are present in multiple domains. For example, flaws in one system can be inherited by its successors and create unintended behavior among interacting systems. In the context of AI, bugs within AI models may be propagated to applications built on top of those models. The financial and cybersecurity case studies further illustrate how risk can arise from banks that are tightly integrated into the global economy and weak linkages in software supply chains, respectively. The case studies also suggest approaches for mitigating systemic risk that can be repurposed for AI.

**Banks**

The financial sector case study highlighted how standards should account for the extent to which one entity's failure might impact the broader system. The Basel Committee on Banking Supervision (BCBS) is a committee hosted and supported by BIS, which seeks cooperation among central banks in pursuit of monetary and financial stability. BCBS establishes standards and best practices for supervision to help BIS achieve its mission.[13] One practice is using an indicator-based measurement approach, where several indicators—in this case bank size, cross-jurisdictional activity, interconnectedness, substitutability, and complexity—are chosen to reflect what makes a bank critical for the stability of the financial system.[14] These indicators inform policy that increases banks' loss absorbency, their ability to sustain losses without falling below a minimum threshold of capital and facing insolvency. Importantly, these indicators can increase the loss absorbency of banks that are of global systemic importance, whose failure could pose a threat to the international financial system in the absence of preventative measures.[15]

FSB, a financial institution that coordinates the work of national financial authorities and international standard-setting bodies, designed a standard that serves a similar purpose to BCBS's measurement approach, but takes a different form.[16] In pursuit of its

singular focus on financial stability, FSB designed the Total Loss-Absorbing Capacity standard, which is defined as the instruments and liabilities that should be readily available for a bank of systemic importance to fail and still provide support to entities around the world. The standard is set as a range of debt percentages, which gives banks flexibility in how they comply with the standard. However, as the presenter for the case study noted, banks typically choose the maximum acceptable debt to assets ratio to maximize profitability despite the higher risk. Therefore, if standards prescribe a range of acceptable values, but businesses have incentives towards maximizing or minimizing that value, standards developers should expect that most businesses will choose that maximum or minimum.

**Cybersecurity**

The cybersecurity case study emphasized the importance of accounting for risks that can emerge from interdependencies between entities in a supply chain. A node in a supply chain can impact its dependencies, posing risk to other nodes within the chain or outside groups that rely on the supply chain. The software supply chain for federal procurement is a critical component of national security, such that Executive Order 14028, Improving the Nation's Cybersecurity, instructed the Director of the NIST and the Secretary of Commerce to issue guidance on enhancing the security of the software supply chain.[17] However, significant challenges for successfully implementing EO 14028 remain. These include the difficulty of standardizing security measures for different software and hardware suppliers, the lack of criteria for what constitutes "trustworthiness" in the software supply chain, and increasing governmental trends towards data localization, which could hinder access to information needed to build software and assess its security.[18]

Widely used AI models can introduce new security risks due to the decentralized nature of software development, as well as the speed and scale of deployment. This is especially apparent for models that are built on open source software. These models are susceptible to malicious updates that target their open source software dependencies. Furthermore, code contributors from around the world can unwittingly introduce vulnerabilities to open source models if they rapidly make changes to a model's software that are not quality controlled.

**Box 1. What is Open Source Software?**

- Open source software is code that anyone can access for free and is often used to build AI models and systems.
- Current AI models have networks of dependencies; for example, large language models hosted on open source platforms such as Hugging Face can be cloned and used in many other settings to perform tasks such as storytelling or summarization.
- Copies of an open source model on local machines can pull updates from the original open source model when needed.

To the extent that open source models have flaws such as security vulnerabilities, these deficiencies may plague dependent systems and proliferate quickly (See Box 1 for more information).[19] Models that are not open source also support downstream applications and may pass their deficiencies on to these applications. For example, GPT-4, a general-purpose model whose technical specifications cannot be accessed by the public, supports applications such as Ask Instacart, EinsteinGPT, and My AI for Snapchat.[20] Flaws that characterize GPT-4 will likely be passed on to these applications, which are deployed in different settings and used by many people. As a result, large numbers of people may be at risk of harm that stems from a single model's vulnerabilities.

A key part of risk management is planning for risks that originate from system interdependencies. NIST has developed an AI Risk Management Framework (AI RMF) that lays the foundation for mapping, measuring, managing, and governing risk from AI systems. It recognizes that harms related to AI systems are of different magnitudes and affect people, organizations, and ecosystems.[21] The Map function of the AI RMF, which establishes the context to frame risks related to AI systems, emphasizes anticipating how these harms might emerge across the AI lifecycle due to interactions between AI actors, or those who play a role in the AI system lifecycle, and systems. These anticipatory exercises can mitigate uncertainty and enhance the integrity of risk management decisions related to AI systems.[22] By linking interdependencies to risk at different scales, the AI RMF demonstrates the applicability of system interdependencies to AI risk management.

## Recommendations

- ***Critical infrastructure owners and operators should track the interdependencies of their AI systems.*** Similar to the way banks of systemic importance impact the stability of the global economy, critical infrastructure operators manage and influence the functioning of vital services. The October 30 Executive Order directs the Secretaries of Homeland Security, Commerce, Sector Risk Management Agencies, and other regulators to incorporate the NIST AI RMF and other appropriate security guidance into relevant safety and security guidelines for use by critical infrastructure owners and operators within 180 days.[23] These guidelines should take inspiration from the financial sector case study and direct organizations that oversee critical infrastructure to develop measurable indicators that reflect the interdependencies of their AI systems, such as their cross-jurisdictional uses and the availability of substitutes. Guidance should encourage organizations to publish these indicators in profiles, or applications of NIST's AI RMF* for different end uses, along with how they integrate these indicators into their AI risk management processes. Interdependency indicators should complement other documentation that describes an AI system's data, algorithms, models, and evaluations. Measurable risk indicators that are articulated in profiles for different end uses, technologies, and sectors will lower costs for other organizations that adhere to the AI RMF since they will need to invest fewer resources in operationalizing AI safety. In the future, NIST or the appropriate agency should develop standard metrics to help networks of organizations across sectors track the interdependencies of their AI systems.

- ***The Office of Management and Budget (OMB) and the Office of Science and Technology Policy (OSTP) Directors' forthcoming guidance on minimum risk management practices for AI should require that agencies identify the risks that could emerge from interdependencies between their AI systems and other entities.*** The October 30 Executive Order directs the OMB Director and the OSTP Director, in consultation with an interagency AI Council that coordinates the development and use of AI in agencies, to issue guidance to agencies that defines minimum risk management practices for AI. The October 31, 2023 OMB draft policy provides more detail about the minimum practices

* CSET has posted high-level guidance and a basic AI RMF profile template to assist organizations creating custom AI RMF profiles.

that agencies must take for AI that impacts rights or safety, stating that agencies should "assess the possible failure modes of the AI and of the broader system, both in isolation and as a result of human users and other likely variables outside the scope of the system itself."[24] The forthcoming guidance from the OMB and OSTP Directors should encourage agencies to expand on how variables outside the scope of their AI systems, and specifically interdependencies between their systems and other entities, can introduce or exacerbate risk. Risk assessments by agencies should include scenarios where forces outside of an AI system, such as its suppliers or downstream applications, create undesirable impacts or are negatively affected. The guidance should also direct agencies to document planned responses to failures of AI systems tightly integrated with other entities, such as hardware devices or additional AI systems, to minimize the likelihood of the system creating harmful ripple effects. Among the considerations should be the criteria under which any fail-safe or override modes are triggered, as well as criteria under which the system should be recalled or shut down.

***Takeaway 2***

***Guidance on testing and re-approval of AI systems should be calibrated to risk and account for changes to AI systems over time.***

Requirements for testing to obtain market approval can be a useful tool to manage the risks of new technologies and are likely to be employed for the governance of AI technologies. These requirements should be calibrated to factors including the degree of risk that a system introduces and how well the system is already understood. Risk-based requirements, where systems that pose higher risks based on their technological characteristics, capabilities, and context of use require more stringent testing, are already employed in some regulatory regimes in the United States and are a prominent feature of the most recent text of the EU AI Act adopted by the European Parliament. The history of medical device testing indicates that while single minor modifications may be permissible without re-testing in limited contexts, larger or more numerous modifications should require at least partial re-testing to guard against unanticipated changes to the safety of systems.

**Medical Devices**

Although the United States is still determining whether and how to classify AI systems according to risk, risk classifications already exist for other sectors and professions, as well as within organizations. The FDA provides a model for handling regulatory approvals for medical devices with different levels of risk.[25] Specifically, it has several different pathways for regulatory approval of medical devices depending on their novelty and degree of risk, which ranges from Class I, the lowest degree of risk, to Class III, the highest degree. Medical devices considered highest-risk, like pacemakers or renal stents, generally support life or are implanted and could cause serious injury if they fail. These devices must always undergo testing before and after the clinical stage, and are subject to ongoing evaluation to ensure their safety and effectiveness, regardless of their similarity to existing devices. For select low and moderate-risk devices, however, there is more regulatory flexibility to strike a balance between safety and speed-to-market.

For devices that are novel, but have a reasonable assurance of safety and effectiveness, the FDA offers an expedited approval process called the De Novo pathway. Typically, a novel device, or a device whose type has previously not been identified or classified, goes through onerous approval processes due to the lack of historical safety and efficacy data. The De Novo pathway, however, allows novel

devices that are considered sufficiently safe and effective to be classified as a Class I or Class II device to increase efficiency. Accelerating such administrative processes when innovations do not pose significant risks can ease the burden of fielding a device and allows useful technologies to be deployed faster.

The approval process can also be accelerated if a device is very similar (in official terms, "substantially equivalent") to a previously approved device (referred to as the "predicate" device).[26] A device is deemed substantially equivalent to a predicate if it has the same intended use and either the same "technological characteristics," including materials, design, energy source, and other device features, or new technological characteristics, but no new safety or efficacy concerns.[27] This approach leverages the fact that, with medical devices, risks are generally traceable to specific, known hazardous technological characteristics—for example, risk of electrocution is introduced by electric current above a specific threshold. If new types of risk are introduced by new technological characteristics, testing is required to reduce uncertainty about their nature in the specific context of the system in use. However, if new technological characteristics introduce no new types of risk, information gathered previously can be reused under the assumption that known risks will not change.

While this process can increase efficiency, it also has unintended consequences. Specifically, devices can be modified a number of times without reapproval, as long as each modification individually does not change the device's technological characteristics, or raise new safety or efficacy concerns. As changes accumulate, a device might be quite different from the original predicate as tested. This is referred to as "predicate creep."[28] In predicate creep, new, unanticipated risks can arise due to the accumulation of interacting technological characteristics, even when each individual change seems benign. Additionally, some devices that pre-date the current FDA approval process have been used as predicate devices despite never having gone through the approval process. According to the presenters, one proposed response to the problems of predicate creep and predicates that have not been explicitly approved is excluding predicate devices that are over 10 years old. At the same time, older devices that have been grandfathered in tend to be simpler, safer, better understood, and less likely to be recalled, so restricting these from being used as predicates may not be ideal.

While the use of predicates may make the regulation of AI systems more efficient, the concept of substantial equivalence is challenging to apply to AI systems for several reasons:

- AI models are often iterated upon gradually (via fine-tuning or online learning, for example) throughout development and deployment, meaning their properties may drift from their predicates and effectively make the predicate a poor point of comparison.
- Limited theoretical understanding of AI means that the effects of changes to aspects of a system such as training data distribution or model size are poorly understood. As systems are modified, risks may increase in likelihood or severity, or entirely new risks may emerge, in unpredictable ways. Therefore, it is difficult to determine whether technical alterations to AI systems create new safety or efficacy concerns.
- Changes in AI systems can occur without any purposeful changes to the system architecture. AI systems that are deemed substantially equivalent to each other before deployment may react differently to the environments that they are deployed in. This may render any previous substantially equivalent designation inaccurate, as similar systems deployed in variable environments can pose vastly different risks over time.

Therefore, the concept of substantial equivalence may only roughly hold for AI systems that are not yet exposed to operational environments.

Although risk-based requirements have not yet been codified in the United States, the White House Office of Science and Technology Policy's (OSTP) Blueprint for an AI Bill of Rights proposes that "high impact risks [should] receive attention and mitigation proportionate with those impacts."[29] Organizations developing or using AI should take heed by calibrating testing of AI systems to their risk and establishing processes for re-testing systems as they change over time.

## Recommendations

- ***The U.S. Department of Defense (DOD) should create thresholds or triggers for different levels of rigor and oversight for testing military AI systems.*** The process of testing systems in the DOD is complex and varied. The rigor, scope, and oversight involved in system testing typically depend upon thresholds related to the system's acquisitions process or cost—or both. For example, systems that cost the DOD at least $3.065 billion in FY 2020 constant dollars to procure are categorized as a Major Defense Acquisition Program. Systems at or above this cost threshold require more thorough DOD testing procedures, with more test planning, resources, and oversight. This approach to modulating test requirements is not unlike the FDA's approach to adjusting test requirements for medical devices based on their risk. Acquisitions processes and cost have historically been sufficient proxies for system importance to DOD missions. However, these are not good proxies for the relative importance or risk of AI systems because low cost, easily acquired AI systems can still cause significant harm by behaving in unexpected ways, such as by providing faulty decision support or inaccurate classification. Thus, DOD should develop new AI-specific thresholds that trigger different testing processes, rigor, and oversight for AI-enabled systems to ensure that these systems support DOD missions as intended.

- ***U.S. government agencies should establish processes for the reassessment and re-testing of systems as they change over time and share these processes with each other.*** Agencies should monitor safety- or rights-impacting AI systems for safety and performance degradation, and conduct human reviews at least annually and after significant changes to the AI system or its context of use.* Agencies must also gradually incorporate new or updated features to prepare for the possibility of adverse outcomes or adversarial attacks, as scaling of compute, data, or model size can sometimes create systems with capabilities that were not predicted in advance.[30] In doing so, agencies should document the circumstances in which changes to an AI system or its operating environment necessitate reassessment or re-testing. After clarifying the number and extent of modifications that can be made to AI

* In its November 3, 2023 draft memorandum for "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence" the Office of Management and Budget (OMB) indicates that this action may become officially required of agencies.

systems, and external events that may trigger re-testing, agencies should share this information with each other. At the very least, agencies can use the processes of other agencies to begin developing their own practices, and at best, agencies can adopt select practices from other agencies that have overlapping or complementary jurisdictions, such as the FDA and the National Institutes of Health, to prevent redundant work. To support this recommendation, NIST and other members of the standards community should conduct research into examining how incremental changes to AI systems affect their safety and efficacy to enable the creation of qualitative or quantitative thresholds for re-approval.

***Takeaway 3***

***Compliance assistance can help small- and medium-sized businesses prepare for and implement AI regulation.***

While standards compliance can be costly and complicated, compliance assistance can help alleviate the burden for companies. Materials that can help companies implement and comply with standards are often behind paywalls, making the process of demonstrating compliance expensive for some companies. This often includes guidance for assessing the quality of standards compliance. For example, BSI Group, the UK's national standards body, charges about $280 per hour to audit the quality management of medical devices and about $480 per hour to review technical documentation.[31] The cost to implement sector-specific standards, including salaries for employees, the reporting process, and systems for implementing compliance, tends to increase with the number of regulations.[32] The use of compliance assistance mechanisms for OSHA standards provides a model for how governments can separate their compliance and enforcement activities—a lesson that may be applied to AI standards.

**Occupational Safety and Health Administration (OSHA)**

Two separate arms of standards implementation handle support and enforcement within OSHA. Compliance Assistance Specialists (CAS) handle educational outreach, whereas the second arm consists of Compliance Safety and Health Officers (CSHOs) who can levy inspection consequences. Compliance Assistance Specialists sit within Regional and Area Offices and promote OSHA's cooperative programs. These specialists offer information about OSHA's compliance assistance resources and conduct informative seminars, workshops, and events, among other functions.[33] On the other hand, CSHOs receive standards implementation guidance from the Directorate of Enforcement Programs and manage enforcement. CSHOs conduct inspections—typically without notice—and issue citations and fines.[34] The separation of these functions enables organizational cooperation with OSHA. Whereas CSHOs bring consequences for non-compliance, compliance assistance measures are managed by different people who engage groups that include small businesses, trade and professional associations, and union locals, helping organizations achieve compliance while implementing OSHA requirements. In Fiscal Year 2022 alone, CAS conducted over 6,200 outreach activities impacting 2.7 million people.[35]

Importantly, the separation of compliance assistance programs can provide businesses that proactively seek to improve their standards implementation with temporary protection from enforcement penalties. The OSHA Consultation Program, for example, is a free, on-site service provided by state governments to help smaller businesses improve their compliance. Consultants do not report violations to enforcement staff and enforcement inspections are suspended until a consultation case closes, which is when corrections must be completed.[36]

Compliance challenges are particularly pertinent for AI, and especially for small- to medium-sized enterprises. Since the development and deployment of AI technologies have so far progressed without much policy intervention, enterprises may have to make significant changes to their workflows depending on the scope of future regulations. Hence, compliance will cost companies and may shape the prospects of small companies in particular. The EU's Impact Assessment of the Regulation on Artificial Intelligence found that small businesses can expect compliance costs around a couple hundred thousand euros for one high risk AI product that requires a quality management system, if no such system is already in place.[37] Therefore, the services offered by OSHA to help companies achieve compliance provide a good model for mitigating risk while helping businesses achieve their goals. Such programs also support AI adoption and innovation by reducing regulatory and compliance uncertainty, especially for smaller businesses with less resources.

### Recommendation

- ***Congress should create a pilot AI Compliance Assistance Office within the U.S. Department of Commerce, which should later expand to other government agencies.*** Although there are currently no U.S. federal AI regulations via legislation, they could be coming soon.[38] To get ahead of any regulatory hurdles and smooth the transition for companies—particularly for small- to medium-sized businesses—Congress should create an AI Compliance Assistance Office within the U.S. Department of Commerce to begin hiring technical experts, build relationships with companies, and define processes. A potential home for the office is the U.S. Commercial Service, an arm of the International Trade Administration within Commerce that already has strong relationships with American businesses. This would serve as a way to mature the compliance assistance function within Commerce and expand the pilot to other government agencies, which could adopt this model once regulations are finalized. Creating a compliance office that is separate from enforcement would demonstrate U.S. leadership in harnessing the benefits of AI on the world stage,

particularly as other countries such as Canada and entities such as the EU consider stronger regulations. It would also reduce the cost burden on companies that can start anticipating regulations to come.

***Takeaway 4***

***Third-party organizations can remove barriers to standards development, implementation, compliance, and tracking.***

Third-party, non-governmental organizations, including professional organizations and certification organizations, can play important roles in removing barriers to standards development, implementation, compliance, and tracking. Professional organizations can serve as useful intermediaries for information gathering, as well as pool resources from smaller firms for access to often-expensive standards documentation. Non-governmental certification organizations can help conduct standards verification activities to reduce the burden of this work on the government. Private organizations can also often develop standards relatively quickly and flexibly, which can later be incorporated into mandatory standards. The OSHA and sustainability case studies reinforce these lessons, which can be applied to AI.

In one example of the inclusion of private third parties in standards work, OSHA's Consultation Program enables private contractors to prepare reports on possible safety violations.[39] By serving as an intermediary to collect complaints from front-line workers in an industry, they can provide anonymity and protection for whistleblowers. This removes a disincentive for employees to report unsafe work practices and provides valuable feedback that can inform risk reduction efforts in the future.

Professional organizations are third parties that can reduce overhead costs for organizations to access standards. Mandatory federal regulations often incorporate copyrighted, industry-designed private standards as references. This practice is useful because it avoids the duplicative or unnecessary creation of standards by the government for its own use.[40] However, incorporating private standards into regulations does not remove them from private ownership, which creates an additional compliance cost for businesses that cannot see the full text of regulations without purchasing the copyrighted standards, or traveling to Washington, D.C. to read physical copies.[41] Small- and medium-sized businesses can be disproportionately affected by the expenses required to access private standards, which may cost hundreds of dollars.[42] These costs can add up when many standards must be purchased to comply.

To address access challenges for standards and to ease compliance, individuals can join professional membership groups to share the cost. For example, a membership for the American Society of Safety Professionals costs $180 annually and includes benefits such as access to industry standards, a network of safety professionals, and career support.[43] These groups are a win for small companies and individuals that struggle to afford standards by making standards that were previously out of reach easily accessible. Professional membership groups not only help small companies and individuals expand the pool of resources they have access to, but may also foster greater adoption of standards themselves.

**Sustainability**

A significant benefit that private organizations can provide to the standards development process is agility. For example, since LEED standards are developed and managed by the private non-profit U.S. Green Building Council (USGBC) outside regulatory processes, they can be updated quickly, which enables them to keep pace with emerging issues and technologies. When appropriate, privately developed standards can also be incorporated by government, as the LEED standard has been used as a baseline requirement for construction and renovation of government-owned facilities.[44] Likewise, leveraging private sector partner expertise through industry associations may accelerate standards creation and adoption.

Standards verification, another stage in the standards development process, can be complex and burdensome. However, the government can make the process more efficient and effective by involving external actors. When third-party organizations are responsible for the certification of standards implementation, these certifiers should follow a consistent, standardized certification framework. In the case of sustainability standards, national accreditation bodies are responsible for the accreditation of the certifiers who enforce proper standards implementation. As one example, a presenter mentioned that the UK Accreditation Service has accredited several carbon emissions monitoring schemes.[45] Without accreditation bodies, certifiers may be inconsistent in their enforcement of standards, and standards can potentially become less effective as organizations may seek out certification from the most lenient of the certifiers.

Collecting information on harm incidents, reducing barriers to accessing standards, encouraging the private sector to contribute to standards, and overseeing certifiers of standards are also critical activities for enabling safe AI systems. For example, companies' internal information on AI incidents could assist in understanding the landscape of AI harm and studying the effectiveness of standards and regulation.

Information gathered from companies through third-party organizations could be aggregated in a way similar to the AI Incident Database, which is a repository of media reports where AI systems have caused or nearly caused harm.[46] Additionally, the private sector has made substantial contributions to AI standards. Private standard-setting organizations such as ISO have published dozens of standards that have been mapped to, and will arguably influence, regulatory and nonregulatory AI frameworks.[47] Finally, certification services for AI are popping up, and there will soon be a need to disentangle the types and quality of certification services that are offered.[48]

### Recommendations

- ***OMB should direct a study by an independent body to inform the designation of third-party accreditation bodies that ensure certifiers evaluate the implementation of AI standards in a consistent manner.*** By default, accreditation should be conducted by existing accreditation bodies, such as the American National Standards Institute National Accreditation Board or the UK Accreditation Service. Alternatively, nongovernmental organizations with technical and standards expertise would be suitable hosts for this function. In the case of AI, this could be accomplished by professional associations for engineering and computing, which are already involved in many standards development processes. OMB should direct a study by an independent body not involved in certification, such as a Federally Funded Research and Development Center (FFRDC), to conduct research into how to most effectively use third-party accreditation bodies to promote consistent AI standards implementation. As emerging and prospective risks characteristic of general-purpose AI systems are now just being explored, particular care should be taken to ensure that accreditation bodies have appropriate incentives and capabilities to evaluate the implementation of standards for AI systems. The study could support effective implementation of the October 30, 2023 Executive Order, which requires the OMB director and other agencies to develop recommendations for evaluation of U.S. government vendors' claims about their AI offerings, and the OMB draft guidance, which requires an independent evaluation authority to review agencies' documentation of consequential AI systems to check that they work as expected and that their benefits outweigh their risks. The study could inform best practices for ensuring that evaluations of vendor claims are performed in a consistent and fair manner, and could separately be used to assist independent evaluation authorities in standardizing their review of AI systems' impacts.

- ***Professional organizations should establish AI standards access funds, whistleblower protection programs, and reporting programs to gather anonymized information on AI risks from industry participants.*** Access to and implementation of standards can help mitigate AI risk, but the cost of purchasing private standards or certification can be prohibitive for certain businesses. Professional organizations should establish AI standards access funds for small- to medium-sized businesses that cannot afford to access standards behind a paywall. Separately, professional organizations should establish whistleblower protection programs to ensure employees are not exposed to undue risk from reporting AI standards compliance violations. Reporting programs for AI risks should complement whistleblower protections. Reports submitted to professional organizations through information gathering initiatives should not be traceable to individual companies so that companies are willing to have their employees participate without reputational risk. The findings of such programs should be shared, at least in summary form, with industry and government stakeholders to inform risk mitigation measures and standards development. These findings could help identify best practices for developing, deploying, and using AI systems and point towards areas where stronger oversight is needed.

***Takeaway 5***

***Non-regulatory governance is one mechanism that can support the safe development and use of AI systems.***

As the financial, cybersecurity, and sustainability case studies revealed, not every form of governance—which includes standards—must be regulatory in nature. Effective non-regulatory methods include:

- Leveraging private and voluntary standards, which build accountability and trust between the public and private sectors;
- Publicly "naming and shaming" entities for failing to uphold standards; and,
- Standardizing government procurement of technologies.

These lessons can inform non-regulatory governance of AI.

While Congress is currently considering more comprehensive AI governance, such as a proposal to create an independent AI oversight body, these plans have yet to become concrete.[49] As these conversations move forward, the U.S. government can glean lessons from the financial, cybersecurity, and sustainability case studies to ensure it upholds a fundamental commitment to a "rules-based and private sector-led approach to standards development."[50] Organizations that have built highly capable AI models have largely been from the private sector, so their technical expertise and perspective should factor into standard-setting discussions.

**Sustainability**

The role of standards put forth by private actors in promoting good governance is underscored by the sustainability case study. The private non-profit USGBC develops and manages standards for LEED, which "provides a framework for healthy, efficient, carbon and cost-saving green buildings."[51] In its role as a non-profit organization, USGBC has shepherded LEED standards to become the model for achieving building sustainability goals. Although there was no regulatory requirement to implement the LEED standard when it was established, it was adopted by real estate developers as a way to demonstrate that a building meets high standards for sustainability. Indeed, researchers have documented the environmental benefit of building to LEED standards relative to conventionally constructing buildings, finding that LEED-compliant buildings contributed 50% fewer greenhouse gases than conventional buildings due to water consumption, 48% fewer greenhouse gases due to solid waste, and 5% fewer

greenhouse gases due to transportation.[52] Moreover, as the presenter at our workshop noted, adoption of LEED standards was also driven by financial motives, as resource-efficient buildings accrue cost savings over time.

**Cybersecurity**

The cybersecurity case study provides lessons about how voluntary standards can help form the foundation for non-regulatory governance. Specifically, this case study highlighted the NIST Cybersecurity Framework, a guidance document that has facilitated the field's maturation and was created through an open, collaborative effort with industry, academia, and government. The initial version released in February 2014 focused on critical infrastructure to fulfill Executive Order 13636, Improving Critical Infrastructure Cybersecurity. However, the framework ultimately provided the basis for broader cybersecurity standardization.[53] This is confirmed by the April 2018 version, which NIST states "evolved to be even more informative, useful, and inclusive for all kinds of organizations" while remaining "flexible, voluntary, and cost-effective."[54] Importantly, the revised framework bolsters a risk-based approach for managing cybersecurity within the supply chain, which fosters public trust in private sector vendors. NIST, the same organization that developed the Cybersecurity Framework, released its first iteration of the AI RMF in January of 2023. Although voluntary consensus standards are still in development for AI, the AI RMF can energize conversations about moving risk management standards forward.

Beyond the Cybersecurity Framework, NIST plays an influential, non-regulatory role in shaping good governance that can be applied to AI standards development. Since the 2018 update, the Cybersecurity Framework has influenced broader standardization language for specific concepts, such as cybersecurity education and workforce requirements, which have their own voluntary resources for businesses.[55] Additionally, NIST has built other non-regulatory functions to keep pace with changes in cybersecurity. A noteworthy example is the Information Technology Lab, which aims to build trust in information technology and metrology by collaborating with industry and other agencies in voluntary consensus Standards Development Organizations (SDOs).[56] Information generation and sharing that occurs within SDOs, as well as throughout the standards development process during public workshops, can make governance activities easier and more achievable for diverse organizations.

**Banks**

The Financial Action Task Force (FATF), an intergovernmental body that sets international standards to counter global money laundering and terrorist financing, similarly relies on information sharing, where the primary goal of sharing is to incentivize countries to change their behavior. The FATF uses "black" and "grey" lists to identify countries with weak enforcement of standards, and then names these countries in public documents three times a year.[57] The black list serves as a call to action for FATF members to apply enhanced due diligence, and sometimes countermeasures, against those countries considered high risk. The grey list names countries subject to increased monitoring as they address deficiencies. These non-regulatory lists have proven effective for protecting the international financial system. By publicly pressuring countries to reform their financial systems, the FATF has seen 72 of 98 black and grey list countries improve their regimes and removed from these lists.[58]

*A "naming and shaming" process could incentivize AI companies to invest in responsible development upfront to avert reputational damage from public censure.*

A "naming and shaming" process, like FATF's grey and black lists, could incentivize AI companies to invest in responsible development upfront to avert reputational damage from public censure. More broadly, an institution akin to a FATF for AI could help heighten public awareness about decisions regarding AI procurement or use by drawing clear attention to actors who are not in line with best practices. "Naming and shaming" is an example showing that non-regulatory governance can be one impactful mechanism for supporting the safe development and use of AI systems.

## Recommendations

- ***The United States should commence discussions in the G7 about creating the equivalent of a FATF for AI.*** Generative AI—particularly ChatGPT—has spurred new fears across the globe about how AI will impact society. Assuming there are more consequential developments on the horizon, a structure akin to the FATF that leverages grey and black lists for AI would place international pressure on countries and other entities to create and implement guardrails for the safe development, deployment, and use of AI systems. Grey lists should include companies and countries that have engaged in questionable conduct around AI. Entities on this list should be subject to increased monitoring and scrutiny. Black lists should include companies and countries that have a documented history of unsafe behavior and show resistance to changing their behavior. Entities that routinely deploy AI-enabled products that cause harm should be added to the black list.

- ***NIST should create an online portal to ensure technical developments relevant to standards are captured and publicized.*** Although the Cybersecurity Framework has been an effective and participatory model for mitigating risks, the updates to the initial framework—which took four years to finalize—are too slow for AI. To ensure the AI RMF remains flexible and relevant, NIST should create an online portal within the existing Trustworthy and Responsible AI Resource Center as an unofficial addendum to the AI RMF, where industry stakeholders can provide real-time updates of AI advancements, such as substantial increases in capabilities of AI systems or decreases in resources required for given capabilities. While NIST should provide oversight of submissions for quality purposes, this portal should serve as an accessible mechanism for entities that may encounter new problems outside the scope of the existing guidance. Further, it would encourage more information sharing among stakeholders. Since the portal would provide an opportunity for prototyping AI standardization language, it would serve as a resource for NIST when the time comes to formally update the AI RMF, which would potentially shorten the revision timeline.

***Takeaway 6***

***Coordination and regular efficacy checks of standards can ensure that standards development is efficient and effective.***

Across sectors, standards development can be a long, multistakeholder process that unfolds over several stages. The OSHA, sustainability, and cybersecurity case studies demonstrated how standards development requires vigilance over time. Standards can quickly become outdated or may hurt disadvantaged groups, so strategies to bolster the effectiveness and longevity of standards are greatly needed. Enabling diverse stakeholders to contribute to standard-setting, and convening these stakeholders regularly to make updates, are mechanisms that help protect the integrity of standards over their lifetime.

Standards development is a multi-year process that should be revisited regularly. Based on a study by the Government Accountability Office, between 1981 and 2010, it took OSHA on average over seven years to develop and issue safety and health standards.[59] Simply identifying the need for a new standard can take upwards of three years.[60] The presenter for the cybersecurity case study reinforced that standards development timelines can be long, noting how the maturation of cybersecurity standards has taken decades.

Even in the post-development stages, standards are not guaranteed to achieve policy goals. An overly lenient standard might be broadly adopted, but fail to significantly improve outcomes in individual cases or in aggregate. On the other hand, an excessively restrictive standard may fail to achieve significant adoption at all, substantially limiting its potential impact on outcomes. For example, the LEED standard has been criticized for not considering the local requirements and priorities of green construction projects in developing countries.[61] Policymakers must carefully consider both the benefits and costs of enforcing a standard.

Standards that take many years to finalize and turn out to be ineffective impose costs on society. New harms that could have been prevented may materialize during the time that the standard is developed and adopted. Upon release, a standard could quickly become outdated in a rapidly changing world. Therefore, standards that are built over a long period of time may need to incorporate new findings that occur during development and be updated after publication.

One strategy for creating effective standards that endure over time is to facilitate coordination among a diverse group of stakeholders. International bodies, nonprofits, and a variety of other organizations that exchange ideas can help infuse different risk perspectives into standards, ensuring that standards are resilient to technological developments. Coordination can take the form of voting. ISO standards, for example, are open for comment every five years, after which changes are voted on by 165 member countries.[62] The sustainability standard-setting bodies USGBC and the Forest Stewardship Council also collaborate on standards because of their standards' interdependencies: while the use of wood in construction can reduce carbon emissions, it may only do so if forests are managed sustainably and the timber is not transported over excessively long distances.[63] However, facilitating coordination among stakeholders can lengthen standards development because aggregating feedback from different groups tends to be time-consuming. Therefore, a balance should be struck between committing to reasonable timelines for standards development and coordinating with as many relevant stakeholders as possible.

Multi-stakeholder coordination can bolster the resilience of standards to emerging risks, but standards should also be monitored and iterated upon to improve their effectiveness over time. Especially with a new technology or new domain, the first version of a standard is likely to be imperfect. OSHA standards, for example, are almost always contested in court after publication by workers or businesses for having overly lax or excessively stringent requirements. As more information becomes available, standards developers and other stakeholders should analyze the effectiveness of the standard and make adjustments as necessary. For instance, researchers have found that LEED standards may not be as tightly linked to environmental outcomes as previously thought, and as noted earlier, LEED standards have been criticized for not being sensitive to emerging markets.[64] Even so, alternative standards that are more achievable for developing markets have filled this niche.[65] This demonstrates how tracking the impact of standards can help address their limitations.

Standard-setting bodies should seek feedback on the impact of standards on downstream outcomes in order to improve them over multiple iterations, but monitoring and updating of standards is often complicated by the difficulty of evaluating their effectiveness. In the case of cybersecurity, for instance, the constant evolution of threat actors' tactics, techniques, and procedures makes it difficult to quantify the effectiveness of specific measures that defend against threat actor behaviors. Similarly, analysis of the counterfactual impact of the LEED standard is complicated by factors including commercial incentives for resource efficiency and pre-

existing public pressure for sustainability. Nevertheless, a helpful first step would be to assess the impact of standards using relatively straightforward measures, such as comparing the frequency of cybersecurity incidents among organizations, and then continue improving the measures as more information is collected.

Since discussions on AI standards are still in the early stages and technical capabilities evolve quickly, the development of comprehensive AI standards may take several years. It is necessary that standards, as well as the regimes that create them, adapt to address the speed of innovation for evolving technologies. While variable timelines for AI standards development, promulgation, and adoption can limit the flexibility of standards, coordination and frequent efficacy checks can ensure that AI standards remain relevant.

### Recommendations

- ***Standard-setting bodies should host biannual summits to coordinate on standards interoperability and efficacy.*** If poorly constructed, incentive structures within standards can conflict and lead to variable outcomes. On the other hand, if incentives are aligned among standards, they can have a powerful effect on organizations building or using AI. For example, sector-specific regulators may establish different review processes for AI systems that are calibrated to their risk. However, these standards could impose conflicting requirements for applications that span multiple sectors, such as general-purpose AI systems. By harmonizing review processes during biannual gatherings, regulators may be able to create multiple layers of protection against harms from general-purpose AI systems. Hosting regular gatherings will also sustain progress on standards for AI evaluation metrics and safety testing, supporting the goals of concordant events such as future international AI safety summits, like the follow on summits to the UK's November 2023 Bletchley Summit.

- ***NIST should support the development of testbeds to monitor AI standards for effectiveness.*** The CHIPS and Science Act of 2022 authorized NIST to create testbeds for "safe and trustworthy artificial intelligence and data science."[66] NIST should support the development of these testbeds and adapt them to function like regulatory sandboxes, but for voluntary standards. Organizations that build AI systems could determine whether their systems adequately implement standards in the testbed with help from subject matter experts. This way, organizations could begin to improve interoperability and demonstrate the

viability of their systems. In turn, NIST could proactively monitor the ease with which organizations adhere to standards, assess if technology has surpassed the scope of existing standards, and vet third-party proposals for additional metrics, benchmarks, and standards. Testbeds for standards implementation would complement the testbeds created by the Secretary of Energy and NSF under the October 30, 2023 Executive Order to advance the safe development of AI technologies.

## Conclusion

The global economy is governed by standards that provide valuable lessons for AI. Through five case studies on standard-setting, adoption, and enforcement procedures from established sectors, we identified a number of insights that should be applied to AI standards. Our discussions highlighted the need to not only address the technical details of standards development, but also the processes. Our recommendations include building structures for standards compliance assistance, leveraging third parties as enablers of standards, incorporating several non-regulatory approaches into the AI governance toolbox, and verifying the efficacy of established standards. Standards should be further developed with an eye to the context of AI systems, including their interdependencies, the degree of risk they entail, and uncertainties based on lack of historical performance data or evolution of systems over time.

A well-resourced, dedicated, and educated workforce is needed to develop and implement such standards. In particular, the cybersecurity case study highlighted how the success of the MITRE ATT&CK® framework for cyber incident reporting largely rested on the capabilities of trained cybersecurity analysts. People and organizations who lead standard-setting activities may require financial support and access to upskilling opportunities, and barriers to recruiting people with different backgrounds and qualifications will need to be lowered. While these considerations are not directly addressed in the report, they are nonetheless pivotal to implementing the report's recommendations. The United States has an opportunity to chart an influential path in AI standards, but only if an informed workforce is at the helm.

*The United States has an opportunity to chart an influential path in AI standards, but only if an informed workforce is at the helm.*

While the U.S. pursues its own AI standards, it must stay attuned to standards produced by its allies and partners. This will help organizations building AI to navigate different governance structures that are emerging globally and prevent a fragmented standards regime. However, the United States must act now to ensure that standards keep pace with the opportunities and risks of AI. Creating standards for AI will not be easy, but examining key lessons for standards development from other industries can make this formidable task easier. The United States should embrace the challenge that lies ahead by leveraging its talent and creativity to repurpose these lessons for AI.

## Authors

Mina Narayanan is a research analyst at CSET working on AI Assessment.

Alexandra Seymour was previously an Associate Fellow with the Technology and National Security Program at CNAS. She completed work on this workshop report in May 2023.

Heather Frase is a Senior Fellow at CSET and leads the AI Assessment line of research.

Karson Elmgren previously worked at CSET as a research analyst focused on AI Assessment.

## Acknowledgments

The authors thank Steph Batalis, Jacob Feldgoise, Mia Hoffmann, Jessica Ji, and Hannah Kelley for their thoughtful feedback on the paper. The authors are also grateful to Margarita Konaev for her careful review of the paper.

The authors thank all workshop participants for sharing their insights during the workshop, and greatly value the feedback from participants who reviewed the paper. Several participants that the authors could acknowledge are listed below in alphabetical order by last name:

- Michael D. Garris
- Chuck Howell
- Sunmin Kim (Applied Intuition)
- Lance Lantier
- Richard Mallah (Principal AI Safety Strategist, Future of Life Institute)
- Deborah Morgan (PhD researcher, Department of Computer Science, University of Bath)
- Carol J. Smith
- Abdul B. Subhani
- Kush R. Varshney (Distinguished Research Scientist and Manager, IBM Research)
- Alexa Wehsener (Deputy Director for Defense Strategy & Research, Institute for Security and Technology)

The authors also thank the case study presenters for preparing presentations and sharing their subject matter expertise at the workshop. Several case study presenters that the authors could acknowledge are listed below in alphabetical order by last name:

- Alexandra Barrage (Partner, Financial Services, Davis Wright Tremaine LLP)
- Keith Bryan (Director, Built Environment, Americas, BSI)
- Sanket Dhruva (MD, MHS, University of California, San Francisco)
- Brendan O'Leary (Former Deputy Director of the Digital Health Center of Excellence, U.S. Food & Drug Administration)
- Kirk M. Sander (Chief of Staff & Vice President, Safety and Standards, National Waste & Recycling Association)



CSET Product ID: 20230021
Document Identifier: doi: 10.51593/20230021
Document Last Modified: 29 November 2023

## Endnotes

[1] The White House, *United States Government National Standards Strategy for Critical and Emerging Technology* (Washington, DC: The White House, 2023), https://www.whitehouse.gov/wp-content/uploads/2023/05/US-Gov-National-Standards-Strategy-2023.pdf.

[2] National Institute of Standards and Technology, "Standards & Measurements," Department of Commerce, June 22, 2015, https://www.nist.gov/services-resources/standards-and-measurements.

[3] "Benefits of standards," International Organization for Standardization, https://www.iso.org/benefits-of-standards.html.

[4] *Conformity assessment – Vocabulary and general principles*, ISO/IEC 17000:2020 (International Organization for Standardization/International Electrotechnical Commission, 2020), https://www.iso.org/obp/ui/#iso:std:iso-iec:17000:ed-2:v2:en.

[5] National Institute of Standards and Technology, "Documentary Standards," Department of Commerce, April 28, 2022, https://www.nist.gov/feature-stories/why-you-need-standards/documentary-standards.

[6] National Institute of Standards and Technology, "Measurement Standards," Department of Commerce, April 28, 2022, https://www.nist.gov/feature-stories/why-you-need-standards/measurement-standards.

[7] Kate Jones, Marjorie Buchser, and Jon Wallace, "Challenges of AI," *Chatham House*, March 22, 2022, https://www.chathamhouse.org/2022/03/challenges-ai.

[8] "Standards catalogue ICS," International Organization for Standardization, https://www.iso.org/standards-catalogue/browse-by-ics.html.

[9] "About OSHA," Occupational Safety and Health Administration, https://www.osha.gov/aboutosha.

[10] "OSHA State Plans: How Many States Have Their Own?" *OSHA.com 360 Blog,* https://www.osha.com/blog/state-plans#:~:text=Most%20OSHA%20State%20Plans%20adopt,hazards%20that%20affect%20their%20workers.

[11] "Rulemaking Process," Occupational Safety and Health Administration, https://www.osha.gov/sites/default/files/rulemaking_process.pdf.

[12] "Information Technology Laboratory Computer Security Resource Center," National Institute of Standards and Technology, https://csrc.nist.gov/glossary/term/risk.

[13] "The Basel Committee – overview," Bank for International Settlements, https://www.bis.org/bcbs/.

[14] Basel Committee on Banking Supervision, *Global systemically important banks: revised assessment methodology and the higher loss absorbency requirement* (Basel, Switzerland: Bank for International Settlements, July 2018), 4, https://www.bis.org/bcbs/publ/d445.pdf.

[15] Office of Financial Research, "Bank Systemic Risk Monitor," Department of the Treasury, https://www.financialresearch.gov/bank-systemic-risk-monitor/.

[16] Financial Stability Board, *About the FSB* (Basel, Switzerland: Financial Stability Board, November 2020), https://www.fsb.org/about/.

[17] Exec. Order No. 14028, 86 FR 26633 (2021).

[18] Erol Yayboke , Carolina G. Ramos , and Lindsey R. Sheppard, "The Real National Security Concerns over Data Localization," *Center for Strategic & International Studies*, July 23, 2021, https://www.csis.org/analysis/real-national-security-concerns-over-data-localization.

[19] Zoë Brammer, Silas Cutler, Marc Rogers, Megan Stifel, "Castles Built on Sand: Towards Securing the Open-Source Software Ecosystem," (The Institute for Security and Technology, 2023), https://securityandtechnology.org/wp-content/uploads/2023/04/Castles-Built-on-Sand.pdf.

[20] Ecosystem Graphs, Center for Research on Foundation Models, https://crfm.stanford.edu/ecosystem-graphs/index.html?mode=graph.

[21] The National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (Washington, D.C.: Department of Commerce, 2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[22] The National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*.

[23] "Executive Order 14110 of October 30, 2023, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." *Federal Register* 88 FR 75191 (2023)*:* 75191-75226. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[24] Office of Management and Budget. *Proposed Memorandum for the Heads of Executive Departments and Agencies on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence.* October 31, 2023. https://ai.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-Public-Comment.pdf.

[25] U.S. Food and Drug Administration, "Step 3: Pathway to Approval," Department of Health and Human Services, February 9, 2018, https://www.fda.gov/patients/device-development-process/step-3-pathway-approval.

[26] U.S. Food and Drug Administration, "Premarket Notification 510(k)," Department of Health and Human Services, October 3, 2022, https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-notification-510k.

[27] "The 510(k) Program: Evaluating Substantial Equivalence in Premarket Notifications [510(k)] Guidance for Industry and Food and Drug Administration Staff" (Food and Drug Administration, July 28, 2014), 6-7, https://www.fda.gov/media/82395/download.

[28] Arianne Freeman, "Predicate Creep: The Danger of Multiple Predicate Devices," *Annals of Health Law Advance Directive* 23 (2014), https://www.luc.edu/media/lucedu/law/centers/healthlaw/pdfs/advancedirective/pdfs/issue11/11_Freeman_Formatted.pdf.

[29] Office of Science and Technology Policy, "Safe and Effective Systems," The White House, https://www.whitehouse.gov/ostp/ai-bill-of-rights/safe-and-effective-systems-3/.

[30] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama et al., "Emergent Abilities of Large Language Models," arXiv preprint arXiv:2206.07682 (2022), https://arxiv.org/abs/2206.07682.

[31] "Fees for Conformity Assessment Activities (EUR)," *British Standards Institution*, November 2022, https://www.bsigroup.com/globalassets/meddev/localfiles/en-gb/documents/bsi-md-conformity-assessment-services-and-fees-eur-uk-en.pdf.

[32] Will Kenton, "Compliance Cost: What it is, How it Works," *Investopedia*, August 13, 2022, https://www.investopedia.com/terms/c/compliance-cost.asp.

[33] Occupational Safety and Health Administration, "Compliance Assistance Specialists (CAS)," Department of Labor, https://www.osha.gov/complianceassistance/cas.

[34] Occupational Safety and Health Administration, "Occupational Safety and Health Administration (OSHA) Inspections," Department of Labor, https://www.osha.gov/sites/default/files/publications/factsheet-inspections.pdf.

[35] Occupational Safety and Health Administration, "Compliance Assistance Specialists (CAS)."

[36] Occupational Safety and Health Administration, "OSHA Fact Sheet: The OSHA Consultation Program," Department of Labor, https://www.osha.gov/sites/default/files/publications/factsheet-consultations.pdf.

[37] Andrea Renda, Jane Arroyo, Rosanna Fanni, Moritz Laurer, Agnes Sipiczki, Timothy Yeung, George Maridis, et al., *Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe* (Brussels, Belgium: Directorate-General for Communications Networks, Content

and Technology, 2021), 154, https://op.europa.eu/en/publication-detail/-/publication/55538b70-a638-11eb-9585-01aa75ed71a1.

[38] David Shepardson and Richard Cowan, "US needs 'comprehensive legislation' to address AI risks, Schumer says," *Reuters*, June 21, 2023, https://www.reuters.com/world/us/us-needs-comprehensive-legislation-address-ai-schumer-2023-06-21/.

[39] Occupational Safety and Health Administration, "OSHA Fact Sheet: The OSHA Consultation Program."

[40] Office of Management and Budget, *Circular No. A-119 Revised* (Washington, DC: The White House, 1998), https://www.whitehouse.gov/wp-content/uploads/2017/11/Circular-119-1.pdf.

[41] Emily S. Bremer, "On the Cost of Private Standards in Public Law," *U. Kan. L. Rev.* 63, (2014): 286, https://kuscholarworks.ku.edu/bitstream/handle/1808/20310/5-Bremer_FinalPDFSheridan.pdf?sequence=1.

[42] Emily S. Bremer, "On the Cost of Private Standards in Public Law," 286.

[43] "Member Types and Qualifications," *American Society of Safety Professionals*, https://www.assp.org/membership/benefits-qualifications/member-types-and-qualifications#:~:text=Annual%20dues%20are%20%24180%2C%20plus,standards%2C%20career%20support%20and%20more.

[44] General Services Administration, "Greening Federal Buildings," General Services Administration, July 6, 2023, https://www.gsa.gov/climate-action-and-sustainability/greening-federal-buildings.

[45] United Kingdom Accreditation Service, "Accreditation: Supporting net zero policies," Department for Business, Energy and Industrial Strategy, October 29, 2021, https://www.ukas.com/resources/latest-news/accreditation-supporting-net-zero-policies/.

[46] AI Incident Database, https://incidentdatabase.ai.

[47] Jessica Newman, "A Taxonomy of Trustworthiness for Artificial Intelligence: Connecting Properties of Trustworthiness with Risk Management and the AI Lifecycle" (UC Berkeley Center for Long-Term Cybersecurity, January 2023), 57-59, https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf; Nativi S. and De Nigris S., *AI Watch: AI Standardisation Landscape state of play and link to the EC proposal for an AI regulatory framework* (Brussels, Belgium: Joint Research Centre, 2021), 20-21, https://www.standict.eu/sites/default/files/2021-07/jrc125952_ai_watch_task_9_standardization_activity_mapping_v5.1%281%29.pdf.

[48] "IEEE CertifAIEd: The Mark of AI Ethics," *IEEE*, https://engagestandards.ieee.org/ieeecertifaied.html.

[49] Senators Richard Blumenthal and Josh Hawley, "Bipartisan Framework for U.S. AI Act," Subcommittee on Privacy, Technology, and the Law, September 7, 2023, https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf.

[50] The White House, *United States Government National Standards Strategy for Critical and Emerging Technology*, 3.

[51] "LEED rating system," *U.S. Green Building Council*, https://www.usgbc.org/leed.

[52] Louise Mozingo and Ed Arens, "Quantifying the Comprehensive Greenhouse Gas Co-Benefits of Green Buildings" (UC Berkeley, October 24, 2014), https://escholarship.org/uc/item/935461rm#main.

[53] National Institute of Standards and Technology, "Cybersecurity Framework: History and Creation of the Framework," Department of Commerce, March 16, 2023, https://www.nist.gov/cyberframework/online-learning/history-and-creation-framework.

[54] National Institute of Standards and Technology, "Cybersecurity Framework: Framework Development Archive," Department of Commerce, April 21, 2023, https://www.nist.gov/cyberframework/evolution.

[55] National Institute of Standards and Technology, "Initial Summary Analysis of Responses to the Request for Information (RFI) Evaluating and Improving Cybersecurity Resources: The Cybersecurity Framework and Cybersecurity Supply Chain Risk Management," Department of Commerce, June 3, 2022, https://www.nist.gov/system/files/documents/2022/06/03/NIST-Cybersecurity-RFI-Summary-Analysis-Final.pdf.

[56] National Institute of Standards and Technology, "Information Technology Laboratory," Department of Commerce, https://www.nist.gov/itl; National Institute of Standards and Technology, "ITL Standards Activities," Department of Commerce, February 15, 2018, https://www.nist.gov/itl/standards-activities.

[57] Financial Action Task Force, *"Black and grey" lists* (Paris, France: Financial Action Task Force), https://www.fatf-gafi.org/en/countries/black-and-grey-lists.html#:~:text=The%20FATF%20identifies%20jurisdictions%20with,CFT%20regimes%20has%20proved%20effective.

[58] Financial Action Task Force, *"Black and grey" lists*.

[59] Government Accountability Office, "Workplace Safety and Health: Multiple Challenges Lengthen OSHA's Standard Setting," Government Accountability Office, April 19, 2012, https://www.gao.gov/products/gao-12-330.

[60] "The OSHA Rulemaking Process," Occupational Safety and Health Administration, https://www.osha.gov/sites/default/files/OSHA_FlowChart.pdf.

[61] Ruveyda Komurlu, Asli Pelin Gurgun, and David Arditi, "Evaluation of LEED Requirements for Site Properties in Developing Country-Specific Certification," *Procedia Engineering* 118 (2015): 1169 – 1176, https://www.sciencedirect.com/science/article/pii/S1877705815021153?ref=pdf_download&fr=RR-2&rr=7fa3f5e498948006.

[62] International Organization for Standardization, *Standards in our world* (Geneva, Switzerland: International Organization for Standardization), https://www.iso.org/sites/ConsumersStandards/1_standards.html.

[63] Forest Stewardship Council, *USGBC Members Approve LEED v4* (Bonn, Germany: Forest Stewardship Council, 2013), https://us.fsc.org/en-us/market/green-building/leed-v4.

[64] Fiona Greer, Josh Chittick, Erick Jackson, Jeremy Mack, Mitchel Shortlidge, and Emily Grubert, "Energy and water efficiency in LEED: How well are LEED points linked to climate outcomes?," *Energy and Buildings* 195 (July 2019): 161 – 167, https://www.sciencedirect.com/science/article/abs/pii/S0378778818327282.

[65] Ruveyda Komurlu, Asli Pelin Gurgun, and David Arditi, "Evaluation of LEED Requirements for Site Properties in Developing Country-Specific Certification."

[66] U.S. Senate Committee on Commerce, Science, & Transportation, "CHIPS and Science Act of 2022 Section-by-Section Summary," U.S. Senate, July 29, 2022, 12, https://www.commerce.senate.gov/services/files/1201E1CA-73CB-44BB-ADEB-E69634DA9BB9.