February 2, 2024

Director Laurie E. Locascio
U.S. Department of Commerce
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

RE: Request for Information Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (88 Fed. Reg. 88368)

Dear Director Locascio:

Adobe Inc. ("Adobe") appreciates the opportunity to submit views to the National Institute of Standards and Technology ("NIST") related to its responsibilities under Executive Order ("EO") 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

As a leader in the development and deployment of advanced AI systems and a signatory to the White House voluntary commitments to manage the risks posed by AI, we recognize the need to build robust AI assurance mechanisms to ensure these systems are designed, developed, and deployed in a trustworthy and responsible manner.

We applaud the Biden-Harris Administration for issuing its landmark Executive Order and NIST for publishing this subsequent Request for Information ("RFI").[1] We thank NIST for continuing to develop guidance documents with significant public input and feedback, including the NIST AI Risk Management Framework ("AI RMF").

Below, we offer thoughts on several of the questions posed by NIST in the RFI, and we look forward to engaging further with the Administration on this important topic.

Our response highlights the following themes:

- **Reducing the Risks Posed by Synthetic Content**. Adobe supports Section 4.5 of the EO, as we have long recognized the importance of building trust in this digital age. Adobe believes the development of effective labeling and content authentication standards (i.e. C2PA) and technologies, like Content Credentials, is critical to fulfilling Section 4.5's goal. Adobe urges NIST and the federal government to promote

[1] 88 Fed. Reg. 88368 (Dec. 21, 2023).

transparency around both synthetic and non-synthetic digital content by contributing to and driving the implementation of Content Credentials—open-sourced technology based on an open standard (C2PA) enabling creators to cryptographically attach provenance information to content so that consumers can see the origins and edit history of content online. The federal government has a critical role to play in facilitating widespread adoption of these standards and tools to help increase the American public's trust in the digital content they consume and enhance their digital literacy.

- **Developing Guidelines, Standards, and Best Practices for AI Safety and Security**. As NIST looks to develop a companion resource to the AI RMF for generative AI, the agency should continue to encourage the adoption of a risk-based approach and work with the private sector to leverage its expertise and familiarity with AI solutions. This will provide organizations and companies clear and consistent guidelines for building internal AI governance structures and equip them with standard methodologies to identify and mitigate AI-related harms. To mitigate harmful bias, Adobe further urges public-private sector collaboration to aid the government's development of standardized test datasets, which would foster greater public trust in AI systems. Adobe is also making recommendations around "red teaming" focused on helping all of us better choose where we make AI red team operation investments, encouraging knowledge sharing and talent pipeline development, and helping lower the costs and barrier to entry for these operations.

Background

At Adobe, our mission is to change the world through personalized digital experiences. Since our founding in December 1982, we have continued to pioneer transformative technologies that allow our customers—who range from emerging artists to global brands—to channel their imaginations, unleash their creativity, and power their businesses.

As technology becomes more advanced, we are committed to innovating responsibly, advancing the responsible use of technology for the good of society and ensuring that our technologies drive a positive impact on the environment and in our communities. This means making sure our technologies and the processes we go through to develop them are accountable, responsible, transparent, and inclusive of our customers and communities. It also means considering the broader implications of new technologies and stepping up as leaders on societal issues where Adobe can leverage our technology and expertise to have a unique impact.

Adobe's business is comprised of three cloud-based solutions: Creative Cloud, Document Cloud, and Experience Cloud. Across each cloud and in our products, Adobe is building on a decade-long legacy of AI innovation by leveraging the power of AI to deliver hundreds of intelligent capabilities. For example, in Creative Cloud, AI powers many creative functions including advanced image editing features in Photoshop, making it easier for everyone to tell

their story with simpler and more intuitive tools. In Document Cloud, AI powers features like Liquid Mode, which automatically converts a PDF layout to different formats to fit different screens so it can be easily read on tablets, mobile devices and more. And as part of our Digital Experience offerings, Adobe's customers can use AI-driven features to deliver relevant and meaningful insights and personalized digital experiences to the millions of visitors on their websites.

In March 2023, Adobe launched Firefly, our new family of creative generative AI models designed to be both creator-focused and safe for commercial use.[2] Currently available in Photoshop, Illustrator, Adobe Express, and Adobe Stock. Firefly allows users to channel their creativity in ways they never imagined possible by simply typing in a prompt and generating images in seconds. In the months since its release, over four billion assets have been created using Firefly.

We believe that AI done right will amplify human creativity and capabilities to new levels with deeper insights, accelerated task performance, and improved decision-making ability. As we continue to harness the power of AI, we are committed to developing and deploying AI in line with our AI Ethics principles of accountability, responsibility, and transparency, while considering its broader impact on society.[3]

Adobe's Comments

In response to the questions posed in the Federal Register Notice, we offer the following thoughts for your consideration.

I. Reducing the Risks Posed by Synthetic Content

Thanks to advancements in AI, developing, editing, and distributing content has become more powerful and accessible for consumers. However, the same tools enhanced by the power of AI to make and share legitimate content can also be weaponized to undermine our national security, sow political and cultural discord, or steal artists' intellectual property. Without tools to help guard against these threats, we risk the inability to tell the difference between fact and fiction, what is authentic or not, and what the true provenance is of a piece of synthetic content.

[2] Frederic Lardinois, "Adobe's thoughts on the ethics of AI-generated images (and paying its contributors for them)," *TechCrunch*, March 22, 2023. Adobe's thoughts on the ethics of AI-generated images (and paying its contributors for them)

[3] Adobe Blog, "Seeing is believing: It's time to restore trust in online media," Dana Rao (May 15, 2023).

New standards and tools pioneered by the Coalition for Content Provenance and Authenticity (C2PA)[4] and the Content Authenticity Initiative (CAI)[5] are now providing people with a simple, reliable method of determining the provenance[6] and authenticity of the content they're consuming, enabling us to build greater trust in this increasingly digital age.

Standards, tools, methods and practices for reducing the risks of synthetic content

A. The C2PA and Content Credentials:

Many efforts to combat deepfakes have previously focused on detection. However, detection technologies continue to face significant limitations: the error rates in detecting fake images or classifying real images as fake are high, AI solutions lack the nuance required to understand that all AI editing may not affect the truth of the image being asserted, and they are not scalable to contend with the sheer volume of deepfakes. In one notable study conducted by Facebook, their deepfake detection challenge was only able to spot deepfakes 65.18 percent of the time.[7]
Consequently, rather than attempting to prove what is false, Adobe focused its efforts on proving what is true and on building a standard and technologies that would allow users to understand the *provenance* of a piece of digital content.

That is why in 2019, Adobe founded the CAI with the goal of restoring trust and transparency in digital content. In just four years, the CAI has grown to over 2,500 members across industries, ranging from technology companies like Adobe, NVIDIA, Qualcomm, and Microsoft; gen AI developers like Stability AI; news organizations like the New York Times and the Wall Street Journal; camera companies like Nikon and Leica to academic organizations, non-profits, and more.

In 2020, the CAI joined forces with Project Origin to establish the C2PA, a formal coalition dedicated exclusively to drafting technical standards and specifications as a foundation for universal content provenance. The C2PA is a mutually governed standards development organization (SDO) under the structure of the Linux Foundation's Joint Development

[4] The Coalition for Content Provenance and Authenticity (C2PA) addresses the prevalence of misleading information online through the development of technical standards for certifying the source and history, or "provenance", of media content. https://c2pa.org/

[5] The Content Authenticity Initiative is a group of creators, technologists, journalists, and activists leading the global effort to address digital misinformation and content authenticity. They are focused on promoting and providing an open, cross-industry approach to media transparency. https://contentauthenticity.org/

[6] Provenance refers to the basic trustworthy facts about the origins of a piece of digital content (photo, video, audio file). That is, how a piece of digital content has changed over time, how it has been edited, combined, manipulated, etc.

[7] James Vincent, "Facebook Contest Reveals Deepfake Detection Is Still an 'Unsolved Problem,'" *The Verge*, July 12, 2020, https://www.theverge.com/21289164/facebook-deepfake-detection-challenge-unsolved-problem-ai.

Foundation. The C2PA's work has produced Content Credentials, the specification for which can be found online.

The C2PA and CAI promote the widespread adoption of Content Credentials. Content Credentials are opt-in, cryptographically signed, tamper-evident data that can be embedded into various types of assets (images, videos, audio, documents, etc.) and essentially function as a "nutrition label" for digital content. They enable users to attach information to a piece of content such as their name, the date, and details on the edits that were made. Because all the data inside a credential is digitally signed, it cannot be modified without breaking the signature, so if any tampering does occur along the way, the tampering will be evident, and the public will know to be skeptical. Content Credentials are based on the same technology (cryptographic hashes and signatures) used by secure web sites and secure document signing, including in the U.S. government. In fact, the U.S. Government Publishing Office (GPO) has been leveraging it to securely publish official U.S. government documents for the last 20 years.

Content Credentials can also show whether AI was used, and more importantly, *how* it was used. This information is embedded into a piece of content's metadata and travels with it wherever it goes. In this way, Content Credentials give individuals, governments, or other users a way to show their work and give people a way to see context alongside the content they are consuming. This increased level of transparency helps drive down the risks of synthetic media, enabling citizens to make more informed decisions about whether to trust the content they see online. Importantly, at Adobe, we have promoted this level of transparency by integrating Content Credentials into popular products such as Photoshop, Lightroom and Firefly.

Beyond Adobe, other companies are similarly integrating the Content Credentials technology in their products. For example, Microsoft, another C2PA founding member, recently announced that they are leveraging Content Credentials in all AI-generated images in Bing, "including time and date it was originally created." Truepic (another C2PA founding member) a pioneer in secure camera capture technology, has also introduced Content Credentials into their solutions, enabling users to securely record, verify and authenticate images of human rights violations in Ukraine. The provenance of the captured images is then securely stored in Microsoft's Azure cloud platform, to be used later to demonstrate the authenticity of the atrocities that have occurred. Finally, this past fall, Leica introduced the world's first camera with Content Credentials built-in, delivering authenticity at the point of capture—a significant milestone for the future of photojournalism.[8] This camera is now selling, and Canon, Sony, and Nikon are slated to follow suit, with their own Content Credential-enabled devices in the coming months.[9]

---

[8] https://leica-camera.com/en-US/photography/content-credentials

[9] Tsuyoshi Tamehiro, Keiichi Furkukawa and Yoichiro Hiroi. *Nikon, Sony and Canon fight AI fakes with new camera tech*, "NikkeiAsia," Dec. 30, 2023. https://asia.nikkei.com/Business/Technology/Nikon-Sony-and-Canon-fight-AI-fakes-with-new-camera-tech

The C2PA Standard – How it Works

Content Credentials' underlying technology is based on an open technical standard developed by the C2PA, which can be found online. The goal of the C2PA specification for Content Credentials is to tackle the extraordinary challenge of enabling trust in media in a context of rapidly evolving technology and the democratization of powerful creation and editing techniques. To achieve this goal, the specification is designed to enable global, opt-in adoption of digital provenance techniques by creating a rich ecosystem of digital provenance-enabled applications to meet the needs of a wide range of individuals and organizations. The specification is also designed to meet appropriate security and privacy requirements, as well as human rights considerations.

Content Credentials are comprised of a series of statements that cover areas such as asset creation, authorship, edit actions, capture device details, bindings to content and many other subjects. These statements, called assertions, make up the provenance of a given asset and represent a series of trust signals that can be used by a human to improve their view of an asset's trustworthiness. Assertions are wrapped up with additional information into a digitally signed entity called a claim.

These assertions, claims, and signatures are all bound together into a verifiable unit called a C2PA Manifest by a hardware or software component called a Claim Generator. The set of C2PA Manifests, as stored in the asset's Content Credential, represent its provenance data. The provenance data is embedded into well-defined locations inside of each type of digital content, so that when later edited or manipulated, such tools can add to the provenance. To increase the resiliency of this technique, the C2PA builds security into its designs as they are being developed. As the threat landscape evolves, the C2PA's security design process evolves to meet new and emerging risks as well, and therefore, uses a focused threat modelling process to support development of a strong security and privacy design. The outcomes of this effort directly influence the documentation about explicit threats and security considerations while also facilitating security-oriented thinking throughout the design process.

The C2PA is also working with the International Organization for Standardization's (ISO)[10] TC 171/SC 2 technical committee[11], whose charter includes content authenticity, to establish Content Credentials as an open standard from a recognized, international standards developing organization.

[10] The ISO is an independent, non-governmental international organization that brings together experts to share knowledge and develop voluntary, consensus-based, market relevant International Standards that support innovation and provide solutions to global challenges. https://www.iso.org/home.html

[11] ISO/TC171 is the ISO committee responsible for the standardization of technologies and processes involving capture, indexing, storage, retrieval, distribution and communication, presentation, migration, exchange, preservation, integrity maintenance and disposal in the field of document management applications. https://committee.iso.org/home/tc171#:~:text=ISO%20TC171%20is%20organized%20into,systems%2C%20and%20authenticity%20of%20information.

B. C2PA & Watermarking

There has been much discussion in the public and private sector about using watermarking to identify AI-generated content and to increase transparency. While watermarking is not a wholly sufficient solution on its own, it is important and therefore incorporated into C2PA's open standard. In fact, imperceptible watermarking and cryptographic provenance are complementary and help to ensure objective understanding of where content originates.

However, as AI becomes more prevalent in the content we generate and consume, we believe that digitally signed provenance, which provides essential information about the origin of the content, will be critical to ensure people can consume content within its proper context. By having this provenance technology associated with content, it allows consumers to see the origins and edit history of the content, its source, and whether it is AI created or human created, or, most likely, a mix of both. Watermarking technology provides an additional signal to both machines and humans that a piece of content has provenance attached, which is readily available to anyone. Together, these methods provide robust and tamper evident means of enabling transparency in content. We believe the most success will come if these approaches are based on open standard-based technology, with open-source implementations. This will enable members of the public to verify important information coming from government agencies.

Recommendations

Adobe supports Section 4.5 of the EO, as we have long recognized the importance of building trust in this digital age. The Commerce Department's guidance on labeling and authenticating digital content will be critical to bolstering capabilities for identifying and labeling synthetic content produced by AI systems, as well as non-synthetic content produced by creators. This will also set the stage for OMB's guidance to agencies and strengthen public confidence in the integrity of official U.S. government online communications.

To that end, we recommend the following:

Direct implementation of content provenance technologies across the government. We believe the government should ensure trust in digital content through cryptographically signed provenance *and* watermarking. The Biden-Harris administration should take a leading role and direct the Office of Management and Budget (OMB) to issue guidance directing the implementation of open, voluntary, consensus-based standards, like the one developed by the C2PA, for all publicly posted images, videos, audio, and document content on official government websites and communications. The broad adoption of the C2PA standard would enable the federal government to share trusted resources and information with the American public.

Ensure online platforms maintain and display Content Credentials. Currently, content provenance metadata is removed from major internet platforms, depriving Americans of critical information and context. We recommend the federal government, working with the Congress, require all online platforms to maintain and display any Content Credentials present in digital content on their systems. Creators are already using this technology to establish provenance and it is important to show their work as a way to establish the provenance of their content on the platforms where users consume it. It should be the policy of all democratic governments that if a piece of content has Content Credentials attached, those credentials should not be stripped away.

Establish international norms for content authentication. As part of Vice President Harris' visit to the United Kingdom to deliver a major policy speech on AI, she called for the development and implementation "of international standards to enable the public to effectively identify and trace authentic government-produced digital content and AI-generated or manipulated content, including through digital signatures, watermarking, and other labeling techniques."[12] We recommend that the Biden-Harris administration leverage its convening authority and work through international bodies to which the U.S. is a party to and lead the development of international norms on content authentication, as well as contribute to and encourage adoption of existing open technical standards for digital content provenance

Develop comprehensive educational and media literacy programming. As part of establishing widespread adoption of the technology, the government has a critical role to play in enabling digital literacy, including public safety campaigns to help users – specifically students understand a) that they cannot trust everything they see and hear online and b) that there are tools to help view and verify the provenance of content before they consume it. For this reason, as part of the Adobe-led CAI, we worked collaboratively with education experts to develop free and publicly available Media Literacy Curricula.[13] U.S. Federal Trade Commission, Bureau of Consumer Protection, in consultation with the Department of Education, the Commerce Department, and the Federal Communications Commission, should stand up a program to promote digital literacy and public understanding concerning emerging technologies, such as Artificial Intelligence, that allow for the generation and manipulation of online content. This program should also educate consumers and businesses about the potential harm caused by AI and digital content forgeries. The program should develop tools and information resources (workshops, public safety outreach campaigns) to ensure consumers are better informed about the prevalence of misleading consumer

---

[12] "FACT SHEET: Vice President Harris Announces New U.S. Initiatives to Advance the Safe and Responsible Use of Artificial Intelligence." *The White House*, 1 November 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/11/01/fact-sheet-vice-president-harris-announces-new-u-s-initiatives-to-advance-the-safe-and-responsible-use-of-artificial-intelligence/#:~:text=International%20Norms%20on%20Content%20Authentication,AI-generated%20or%20manipulated%20content%2C

[13] https://edex.adobe.com/cai - the CAI's Media Literacy Curricula were designed to help middle school, high school, and college/university students develop critical media and visual literacy skills to better navigate the ever-changing digital information landscape.

information online and the transparency solutions and standards (i.e. digital content provenance and watermarking) that exist out in the market to help them verify, then trust, content before they consume it.

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

A. Guidelines to Promote Consensus Industry Standards in the Development and Deployment of AI Systems

Adobe supports a risk-based approach to AI in line with the NIST AI Risk Management Framework (AI RMF). As NIST works to develop a companion resource to the AI RMF, the agency should continue to leverage the private sector's expertise and familiarity with AI solutions to develop clear and consistent guidance for internal AI governance. This would better equip companies with standardized methodologies to identify and mitigate AI-related harms. Beginning in 2019, Adobe proactively and voluntarily developed a layered, multi-disciplinary process for responsible AI, which included establishing an AI ethics program with a comprehensive review process that includes an impact assessment. As such, we suggest four elements be considered when establishing an internal AI governance model. We also believe these elements could be incorporated into any future AI RMF companion resource.

1. **Establish a set of AI ethics principles**: In creating our AI governance model, we started with establishing a set of principles – accountability, responsibility, and transparency -- that would guide the development of our AI-powered solutions. As organizations construct their own governance models, there needs to be a "north star" or set of principles from which all other activities flow. This begins with establishing a set of AI ethics principles that align with that company's or organization's mission and philosophy.

2. **Create a comprehensive AI ethics review process and board**: Guided by a set of principles, companies can establish standardized processes for overseeing the design, development, and deployment of ethical AI systems. To ensure the proper governance and oversight of this process, we recommend the creation of an AI review process and establishing an AI ethics review board that can take action to address ethical concerns that arise with new features and technologies. At Adobe, we have implemented similar structures and processes: our AI Ethics Committee is a diverse, cross-functional team with members from across Adobe that is responsible for helping to ensure that the AI Ethics principles are understood and incorporated across our development teams, while our AI Ethics Review Board is tasked with reviewing AI-powered features and products before their release.

   We also recommend organizations consider creating an AI Ethics team whose mission would be to (a) advise product and development teams on ethical considerations for AI-powered solutions and (b) to think expansively about ethics across various types of

tools and technologies that would be used by their customers, and internally for their own use as well.

3. **Establish a risk-based approach to foster innovation**: Establishing a risk-based approach to AI will foster innovation but also require organizations to conduct a more rigorous and thorough review of the high-risk AI systems. At Adobe we have a multi-part review process with an integrated AI ethics assessment that is designed to assess the ethical impact of an AI product or feature. If an initial assessment shows no major ethical impact and the feature meets our ethical standards, the feature is approved. For example, an AI feature that recommends a set of fonts based on a document template would move through a faster review process. AI features that have a higher potential ethical impact go through a more rigorous testing process, including a review by the AI Ethics Review Board.

4. **Ensure Diverse Human Oversight**: Having diverse human oversight and diversity of thought over the lifecycle of the AI feature is important. Diversity matters at every stage of the process. When AI innovations are developed collaboratively with diverse groups of employees, customers, and stakeholders from a range of backgrounds and expertise, the different perspectives naturally surface issues and identify opportunities a more homogonous group would not see.

Further Recommendations

Pursuant to Sections 4.1(a)(i)(A) and (C) of EO 14110, we offer the following technical and governance best practices for implementing trustworthy AI.

Develop Standardized Test Datasets: We believe creating and managing industry verticalized test datasets will both accelerate innovation and foster trust in AI systems. Specifically, companies today use test datasets to ensure their AI models are generating responses against an acceptable range of results. Using a test dataset ensures consistency across different models and different versions of models. For example, a test dataset can be designed to identify bias or harm in an AI model by the type of data in the test dataset. Testing against that test dataset helps the AI developer understand if the AI model meets internal company standards. However, having this be a company-by-company process creates inefficiency and public uncertainty in the quality of the model as every company is using its own acceptable range of results and test datasets. The government is well positioned to accelerate innovation and set the standards for ethical behavior of a model by creating industry vertical-based test datasets that companies could test their models against, certify against a government "range of results" and then ship their product knowing the company met the standard. This provides the companies certainty in releasing models and provides the government oversight on the ethical standards on models that are developed, while avoiding burdensome processes such as requiring companies to disclose billions of training data

assets, or complicated and proprietary testing and development workflows. This approach thus can balance protecting the trade secrets of those who develop and deploy AI models, allow people who use AI models to be equipped with the information necessary to understand the harmful bias risks in the AI feature, and allow consumers to trust the AI innovation.

Continuously evaluate for harm and bias: Internal and external testing, including functional red teaming, before the model release is instrumental to shipping the AI features responsibly. Our internal testers bring diverse perspectives and identities to their experience with a particular feature and feel empowered to shape the features with their feedback. Our approach to task and test set design reflects the AI Ethics principles and the content moderation standards expected from the feature. The evaluations for harm do not stop with a model release but continue for the new model versions adding to the comprehensive model quality evaluation process.

Establish content guidelines for AI-generated content: While we have made significant progress identifying acceptability criteria for AI-generated content that reflect Adobe principles and values, having an industry-wide standard for generated content guidelines, like a movie ranking system, would lead to a greater alignment and easier decision making.

Pre-process data: Making sure that the training data is free of the content that we do not want to see as the AI feature output, like sexual content, Child Sexual Abuse Material (CSAM), or extreme violence, has been a strategic effort for our teams. Data debiasing is another related initiative aimed at removing biased association from the language metadata used in model training. By establishing requirements to pre-process and de-biase data, AI developers could take a meaningful step towards ensuring an AI model's feature does not produce harmful outputs.

Establish guardrails for input and output processing: Our teams focus on implementing guardrails for both user input and model output: the input safety guardrails include both classifiers and lookup lists that ensure the user input does not violate our ethical content policies, and the output guardrails, such as Not Safe For Work (NSFW) classifiers, help prevent unintentional harms of accidental generation of undesirable content. We recommend the NIST encourage the adoption of similar guardrails for processing inputs and outputs.

Create mechanisms for receiving feedback and monitoring AI product features: Monitoring and addressing user feedback is critical for our AI features, and to AI governance in general, as they emphasize feature observability and give voice to our customers. While in-feature feedback channels, such as Report and Flag functionalities, are critical for an AI feature, active engagement with our customers is equally important, and extends to organizing user community events and conducting research focus groups, surveys, and interviews. Establishing some type of mechanism to solicit feedback about an AI feature, or having a process to monitor a feature, would provide AI developers critical information about a

feature's performance. Armed with this information, developers would be able to ensure the product is performing as intended, free from any harm or bias, or take necessary corrective action.

B. Guidelines to Conduct AI Red-Teaming Tests

Functional red teaming to probe for harms, biases, and defenses against potential adversary behaviors is critical for responsibly launching an AI feature. Adobe cautions, however, that red team exercises are currently quite resource intensive operations with long lead times per activity. Our own decisioning on which scenarios qualify for this significant resource investment is determined by comprehensive threat modeling combined with internal and external adversary intelligence that helps narrow the scope of our investment to those of highest attack interest. This high resource cost means that these exercises today are normally only undertaken by larger companies with significant resources like Adobe – and even then, only in the narrow range of threat scenarios that we determine will best benefit from the red teaming approach. A majority of threat scenarios, in our view, can be effectively defended against through other elements of a strong security program including threat modeling, code scanning, internal and 3rd party pen testing, and nurturing a strong community of security researchers through bug bounties and other development programs.

Adobe does believe, however, that there would be significant benefit to both industry and government in NIST providing guidance and programs that help all of us better choose where we make AI red team operation investments, develop a diverse pipeline of talent to help build AI-ready red teams, encourage broader knowledge sharing, and ease the entry point to effective red teaming for AI technology developers with more limited resources. This includes, but is not limited to, the following:

Establish AI threat modeling and incident scoring guidance. Red team operations are uniquely suited to combating specific types of high-value attacks against any type of system – including AI systems. For example, AI technologies themselves have the potential to "supercharge" threat actor activities. Thus, red team operations can be essential in combating those activities. Red teaming also involves creating both realistic scenarios reflecting the intended use while challenging attempts to 'break' the system. To assist in making appropriate resource investment tradeoffs, NIST can provide deeper threat modeling and incident scoring guidance specific to AI systems. This could include a knowledge base of AI-focused comprehensive sample threat models developed in consultation with industry. Such investment would help AI developers make better decisions about where to focus their red teaming efforts.

Expand threat intelligence resources. Expand NIST's adversary intelligence sources to widen available threat intel on AI system attacks to drive red-teaming decisions. Any effort here should be focused on expanding shared knowledge around threat actor tactics, techniques,

and procedures (TTPs) specific to AI systems. This could include, but is not limited to, investment in expansion and use of established sources such as MITRE ATT&CK. Focus and investment here would assist in helping to reduce the lead time for red team operations.

Create shared resources documenting red-teaming best practices. NIST should work with Adobe and other industry partners to document best practices for both starting and developing red teams. This would include recommendations on capabilities, rules of engagement, and how to measure success. Recommendations can also be provided on developing a documentation standard and schema to capture successful and unsuccessful attack simulations including such parameters as tested assumptions (expected feature behaviors), inputs leading to an incident, a received output, risks that an incident represent, an environment (desktop/Web/mobile version), a timestamp, a model version, a severity level, a type of occurrence (initial or repeated), other details (e.g., number of iterations leading to an incident, comments, tested guardrails, suggested mitigations). It should also include comprehensive documentation examples around the TTPs used by red teams, especially when conducting operations for AI-based systems, to help reduce lead time in both scaling resources and operations. Success measurement for red teaming is also a controversial subject in industry today given red team operations generally are "pass" or "fail" operations. Adobe recommends that the definition of success be focused more on measuring maturation of team capabilities, and the ability of the product to withstand a maturing adversary, than any individual operation. NIST can work with industry to develop guidelines based upon established red team maturation models specific to AI readiness, such as the Red Team Capability Maturity Model, and encourage regular assessments. All of these efforts can help make red team operations more effective and repeatable as needed.

Establish a red-team community development program. Most industry knowledge around red teaming is "tribal knowledge," which limits effective sharing. Thus, there are very few community and industry activities focused on red teaming. We do believe there is an opportunity for NIST, working with industry partners, in breaking down these knowledge silos around red teaming. Adobe recommends that NIST replicate its success with peer agencies and industry around security researcher recruitment and development to foster the creation of a red-team community development program that cultivates content, programming, conferences, consortia, and other activities to build a more vibrant red teaming community.

Diversify training, skilling, and recruitment of red-teams. Given its newly increased popularity, many skilled experts on AI security testing are not in the private sector but perform their research in academia. NIST can help encourage information exchange between academia, government, and private sector to allow for the rapid integration of testing techniques into red teaming technique libraries. One of the most important efforts Adobe

believes that NIST, in partnership with CISA and other peer agencies, can leverage to develop more effective red teaming around AI systems is encouraging implementation of the National Cyber Workforce and Education Strategy (NCWES). Successful red teaming at its core is about adoption of a philosophy – the "red team mindset" – combined with the tools and resources necessary to conduct successful operations. A red team should consist of individuals who bring diverse backgrounds, expertise, and viewpoints to the table. A balanced team including both people not actively engaged in the feature development and those with a deep knowledge of the feature has helped in detecting a wide spectrum of issues. We need people willing to adopt this mindset to think of things no one else has thought of, enabling us to test in unexpected ways. Diversity in red teams also helps prevent potential bias – as bias hampers red-teaming operations. This means we must continue to expand the definition of who can do cybersecurity, as championed by the NCWES, beyond the typical requirements for computer science or engineering degrees and untenable amounts of prior experience. Adobe's own successful head of red teaming operations, for example, does not have the background our industry has typically required for these roles. NIST should work with industry partners and peer agencies to develop strong guidelines around the capabilities that will make a good red teamer, aligned with the principles established in the NCWES. It should then encourage and incentivize development of a diverse community of red team members and leaders to help grow the talent pipeline for all of us.

Adobe believes implementing these recommendations will help encourage safer AI systems for all – particularly in helping reduce the barrier to entry for AI technology developers with more constrained resources in both adopting red teaming and actively participating in the broader red teaming community.

Adobe is also committed to continuing to leverage our mature security program to help protect our intellectual property, products, and systems, including AI models and features. We commit to regularly share information, best practices, and learnings from how we test and protect the security of our AI models with the wider community.

Conclusion

Adobe appreciates the opportunity to comment on NIST's responsibilities under the AI EO. We understand and appreciate the efforts you have taken to advance the U.S. government's critical role through the issuance of the EO and this subsequent RFI and look forward to continuing to work with you to promote AI's responsible development and use.