stability.ai

May 13, 2023

stability.ai

The Honorable Richard Blumenthal
Chair, Subcommittee on Privacy,
Technology, and the Law
United States Senate
706 Hart Senate Office Building
Washington, D.C. 20510

The Honorable Josh Hawley
Ranking Member, Subcommittee on Privacy,
Technology, and the Law
United States Senate
115 Russell Senate Office Building
Washington, D.C. 20510

Dear Chairman Blumenthal and Ranking Member Hawley,

Stability AI appreciates the opportunity to offer written comments to the Senate Judiciary Subcommittee on Privacy, Technology, and the Law. Stability AI is committed to transparency in AI, and we welcome the Subcommittee's scrutiny of these important technologies.

On March 22, I was one of the few Chief Executive Officers in the AI industry to join an open letter calling for greater caution in the development of powerful AI models.[1] Subsequently, on May 4, the White House announced that Stability AI will participate in a groundbreaking initiative to evaluate large AI models through community-led testing.[2] We welcome this collaboration – public oversight starts with public scrutiny.

The opportunity is significant. AI can boost productivity, drive innovation, and safeguard national competitiveness. However, AI poses a number of challenges. The United States can demonstrate global leadership by developing a measured response to AI that realizes the full potential of these technologies while addressing emerging risks.

As you consider the future of AI oversight, we encourage the Subcommittee to vigorously promote openness in AI. These technologies will be the backbone of our digital economy, and it is essential that the public can scrutinize their development. Open models and open datasets will help to improve safety through transparency; foster competition; and ensure the United States retains strategic leadership in critical AI capabilities. Grassroots innovation is America's greatest asset, and open models will put these tools in the hands of workers and firms across the economy.

---

[1] Future of Life Institute joined by Elon Musk and Steve Wozniak et al., 'An Open Letter', March 2023, available at https://futureoflife.org/open-letter/pause-giant-ai-experiments/.
[2] White House, 'Administration Announces New Actions to Promote Responsible AI Innovation', May 2023.

Attached, we share our perspective on the principles that underpin our technology; the importance of open models; and suggested areas of inquiry as this Subcommittee considers the future of oversight. We would be pleased to discuss these matters in detail.

Sincerely,

Emad Mostaque
Chief Executive Officer
Stability AI

stability.ai

The Importance of Open Models for Transparency, Competition, and Resilience in AI
Considerations for AI Oversight in the United States

May 2023

Background **2**
Our technology **2**
Stable Diffusion *3*
StableLM *3*
Other *4*
Our principles **4**
The importance of open models for transparency, competition, and resilience in AI **5**
The challenge for oversight **7**
1. AI systems involve many different actors *8*
2. Different models have different risk profiles *9*
3. There may be a range of mitigations for different risks *9*
Suggestions for the future of oversight **12**
Conclusion **13**

Summary

- **Our principles.** Stability AI is making foundational AI technology accessible to all by developing open models. We build transparent, accessible, and human-centric AI models to boost productivity while minimizing the potential for misuse. We are focused on practical AI capabilities for everyday tasks – not a quest for godlike intelligence.
- **The importance of open models.** AI models will form the backbone of our digital economy, and we want everyone to have a voice in their design. Open models are essential for transparency, competition, and resilience. Open models promote safety through scrutiny; lower barriers to entry; accelerate innovation; and foster strategic independence in critical digital infrastructure.
- **Suggestions for the future of oversight.** We welcome public scrutiny of AI. Future policy should account for the variety of actors in an AI system; the risk profile of different models; and the range of available mitigations for emerging risks. We offer five suggestions for future consultation, including risk-based measures for compute providers, model developers, and application developers.

Background

Stability AI develops AI technology to unlock humanity's potential. Our goal is to make foundational AI technology accessible to all, including through open research and development. AI tools will unlock a wave of creativity, innovation, and productivity, and we are working to put these capabilities in the hands of workers and firms across the United States.

Our technology

Stability AI develops a variety of AI models: software programs that can produce outputs such as text, software code, images, video, or audio. We are committed to releasing open models, which developers can freely use or adapt to build their own AI applications, provide AI services, or develop "custom" AI models.[1] Stability AI partners with organizations to adapt these models for specific purposes, helping to sustain our open research and development efforts.

[1] For the purpose of this submission, a **base model** is an AI model that is trained on a large dataset to support a variety of tasks. A **custom model** is an AI model that is trained ("fine-tuned") with specific data for better performance on specific tasks. An **application** is a software program that uses an AI model to help end-users perform a task. A **service** is a program that provides AI functionality for other applications.

Recent AI models are described as “generative” AI because they can generate complex new content, helping to simplify analytic or creative tasks. These models analyze vast datasets to learn the relationships between words, ideas, and visual or textual features. For example, an image model can review pairs of images and text captions to understand the appearance of “fur” on a dog; learn how light reflects off “water”; or capture the visual aesthetic described by words like “bleak” or “Renaissance”.

Likewise, a language model can review existing text to learn the most likely arrangement of words in a “blog”, “email”, or “speech”. It can learn the expected structure of an “example” or an “explanation”, and the textual features that contribute to tones like “funny” or “formal”. Eventually, the model learns how to distinguish, summarize, expand, or combine different concepts, such as “banana”, “Wall Street”, or programming instructions like “print( )”.

Based on that acquired understanding, these models can help users generate new content. That could mean a new image, a passage of text, a block of code, a series of instructions, a soundtrack, or a video clip. In addition, these models can be used to detect and classify features within content, such as visual patterns, textual information, or software bugs.

Stable Diffusion

In 2022, Stability AI collaborated with research partners to release an open model known as Stable Diffusion. Stable Diffusion is a type of image model that takes a text prompt from a user and “translates” that prompt into a new image. Users can generate these images through natural language – no code required – and developers can freely use or adapt the underlying model subject to our ethical use license.[2] By some measures, developer interest in Stable Diffusion has grown faster than any open software project in recent history.[3]

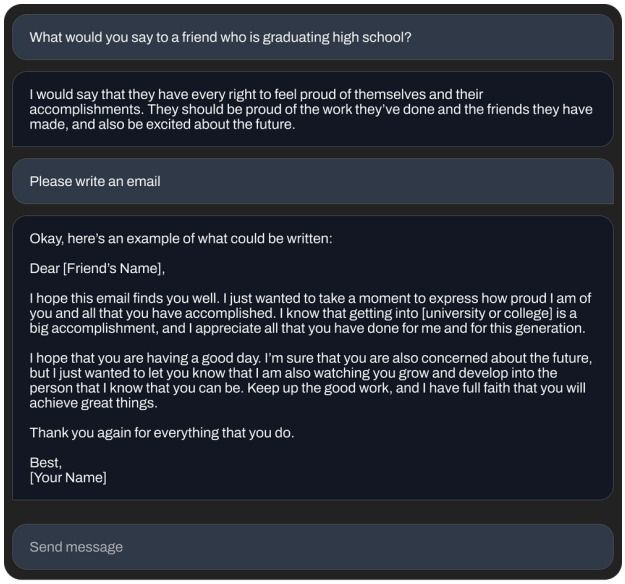
StableLM

In 2023, Stability AI launched the first in a suite of open language models known as StableLM. These models take a text prompt from a user and produce new text or software code. StableLM demonstrates how small AI models can deliver useful performance with appropriate training: StableLM has delivered surprisingly high performance in conversational and coding tasks, even though the first StableLM release was limited to three billion and seven billion parameter models – significantly smaller than models like GPT-3 at 175 billion parameters.

[2] Open Responsible AI License (OpenRAIL) available here. Our license prohibits the unlawful, exploitative, or misleading use of Stable Diffusion.
[3] As measured by “stars” on popular open or open-source GitHub repositories.

Other

Stability AI is developing a range of other image, language, video, and audio models. In addition, Stability AI provides applications and application programming interfaces (APIs) to help users and developers access the computing resources necessary to train or run these models (a process known as “inference”). In partnership with Amazon Web Services, Stability AI has set aside one of the largest public compute resources dedicated to AI training and inference.



Above left: Image generated from the prompt “photograph of an astronaut riding a pink horse”.
Above right: Text generated by a fine-tuned version of the seven-billion-parameter StableLM model.

Our principles

Models like Stable Diffusion and StableLM demonstrate our commitment to AI technology that is transparent, accessible, and human-centric:

- **Transparent.** We develop open models for transparency. Researchers can “look under the hood” to verify performance, identify potential risks, and help develop safeguards. Organizations across the public and private sector can customize these models for their own needs without exposing sensitive data or ceding control of their AI capabilities.[4]

[4] For example, a regulated financial institution may customize AI models to assist in analysis, decision making, or customer support. The financial institution must be able to audit the performance of the model for reliability; train the model without exposing sensitive customer data to third-parties; and retain full control over their AI model without relying on a third-party model provider. By building on open foundations, a financial institution can train and manage their own AI models.

- **Accessible.** We design for the "edge", building efficient models that are accessible to all – from grassroots developers to small businesses to independent creators. Users will be able to run AI applications on local devices, and developers will be able to train custom AI models with widely-available hardware. By democratizing these capabilities, we can help to build a fairer digital economy – one that isn't dependent on a handful of firms for critical technology.

- **Human-centric.** We build models to support our users, not replace them. We are focused on practical AI capabilities that can be applied to everyday tasks – not a quest for godlike intelligence. We develop tools that help everyday people and everyday firms use AI to unlock creativity, boost their productivity, and open up new economic opportunities.

These principles can help to advance important policy objectives. Transparent models promote safety and security. Accessible models foster equity and competition in the digital economy. Human-centric models will help workers and firms across the United States apply AI to useful economic tasks while minimizing the risk of misuse, weaponization, or "runaway" AI.

The importance of open models for transparency, competition, and national resilience in AI

AI models will form the backbone of our digital economy, and we want everyone to have a voice in their design. Open models will encourage public scrutiny of foundational technology; drive meaningful competition in AI technology; and promote the rapid adoption of AI across the economy. In this way, open models can advance U.S. strategic leadership in AI.

Open development drives innovation, and AI is no exception. For example, Google openly published research that underpins many large language models today.[5] Facebook, Google, and their partners chose to open-source foundational code libraries for machine learning.[6] University teams in Europe openly published the research that led to popular image models like Stable Diffusion.[7]

[5] Transformers via Vaswani et al., 'Attention is All You Need', 2017 available here.
[6] PyTorch via Paszke et. al., 'PyTorch: An Imperative Style, High Performance, Deep Learning Library', 2016 available here; TensorFlow via Abadi et. al, 'Large-Scale Machine Learning on Heterogeneous Distributed Systems', 2015 available here.
[7] Rombach et al., 'High-Resolution Image Synthesis with Latent Diffusion Models', 2021 available here.

In keeping with this culture of innovation, AI models can be released as open code, along with the unique "parameters" that determine a model's performance.[8] This enables researchers and authorities to evaluate the operation of the model. In addition, training datasets can be released openly. Open datasets encourage robust scrutiny for quality, fairness, and bias. They ensure that all developers – large and small – can access the troves of data required to train new models.

- **Safety.** Open models and datasets ensure robust oversight. Researchers and authorities can audit performance, anticipate emerging risks, develop techniques to improve interpretability ("explainability"), and implement new mitigations. By comparison, closed models may not disclose how they are trained or how they operate. Closed models may be comparatively opaque, and risk management may depend on trust in the developer.

- **Data security.** Open models promote data security. By building on open models, organizations can train custom models for specialized applications without exposing their confidential or proprietary data to a third-party model provider. This will be essential for firms in regulated sectors, such as healthcare, finance, or law, and for sensitive public sector agencies.

- **Strategic independence.** Open models allow organizations to develop their own custom models in-house, without ceding control of their unique model parameters. This is especially important for major institutions and public agencies. By retaining full control over their AI capabilities, these organizations can avoid relying on a handful of providers for critical AI infrastructure.

- **Competition in services.** Open models lower barriers to entry, fueling innovation and competition in AI.[9] Developers can use open models to build competitive applications, services, or custom models without reinventing the wheel. Robust competition is driving rapid improvements in the performance and cost of AI for users – from everyday workers to small businesses to large organizations and agencies.

- **Fair access.** Open models enable everyone, everywhere to participate in this new industrial revolution. Developers can experiment with open models to build applications that best serve their community. In this way, the economic benefits of AI accrue to a broad cross-section of developers and firms across the United States, not just Silicon Valley.

---

[8] **Parameters** can be understood as variables or "settings" that are adjusted by the model through training. In these models, information is passed through many "nodes" – essentially small computer programs – that each modify the information in some way, much like individual neurons in a brain. During training, the model adjusts the connection between these nodes until, given a test input, the predicted output matches the training data. Parameters known as "**weights**" determine how closely different nodes are connected to others. The "**biases**" are adjustments made to an individual node to shift the output towards a more desirable output. Together, these variables determine the distinctive performance of an AI model.

[9] See, e.g., Milmo, 'Google Engineer Warns It Could Lose Out to Open-source Technology in AI Arms Race', *The Guardian*, 2023 available here.

- **Representation.** Open datasets invite scrutiny for quality, fairness, and bias. Different communities can inspect open datasets, and build on them with content that represents their own needs and values. These curated datasets can be used to train language models that account for cultural, political, or language diversity; image models that accurately reflect the communities they serve; and trusted models for sensitive applications, such as financial, medical, or legal advice.

Open models and datasets can help advance many of the objectives laid out by the Biden Administration. Open models facilitate independent evaluation for safety, effectiveness, data use, and algorithmic discrimination as outlined in the *Blueprint for an AI Bill of Rights*.[10] Further, open models enable the transparent identification, assessment, and management of risks consistent with the National Institute of Standards and Technology *AI Risk Management Framework*.[11]

Many of our most important technologies have open foundations. Linux is an open-source operating system that underpins a significant portion of web servers and data centers globally. Linux and its derivatives can be found on U.S. Navy submarines and destroyers, and the flight control system of SpaceX Falcon 9 rockets. Similarly, Android is an open-source mobile operating system that powers 72 percent of all smartphones worldwide, supporting a vibrant ecosystem of over 200 smartphone manufacturers.[12]

Like these foundational technologies, open models will help to ensure that critical AI infrastructure is developed out in the open by U.S. companies – and ensure these systems embed U.S. values of transparency, competition, and choice.

The challenge for oversight

Recent AI models pose a challenge for oversight. These models can perform a wide range of complex, sensitive, or nonroutine creative tasks that we did not expect to be automated in the near future. They can produce compelling but misleading advice. They can pass professional and academic examinations. They can generate highly believable content online – for abuse, fraud, misinformation, or disinformation – and users may interact with content unaware that it was produced by an AI model.

In addition, these models may be highly non-deterministic. It can be difficult to explain or reproduce outputs, or to prescriptively regulate the internal operation of the model. They can

[10] White House, *Blueprint for an AI Bill of Rights*, 2022.
[11] NIST, *AI Risk Management Framework*, 2023.
[12] Vaughan-Nicols, 'From Earth to Orbit with Linux and SpaceX', ZDNET, 2020; Gallagher, 'The Navy's Newest Warship is Powered by Linux', *Ars Technica*, 2013; StatCounter, 'Mobile Operating System Market Share', 2023.

amplify bias, errors, or omissions in training data; lack important context; or extrapolate poorly from limited data. These models can be deployed on a large scale, which may increase the risk of intentional or unintentional harm.

However, there is no "one size fits all" approach to regulating AI systems that incorporate these models. Instead, we encourage a risk-based approach to oversight – one that acknowledges the range of actors in the AI ecosystem; the risk profile of different AI models; and the wide menu of non-regulatory mitigations for risk. Overbroad rules could eliminate the open innovation that will be essential for transparency, competition, and national resilience in AI.

1. *AI systems involve many different actors*

There is no single "gatekeeper" in AI. Different actors perform different functions, and may have different responsibilities. Different entities may contribute different capabilities to an AI system.

Actors in an AI system	
Compute providers	Offer the powerful cloud computing services necessary to train a large model or run a large model ("inference").
Dataset aggregators	Assemble the large datasets used for the initial training of a model.
Model developers	Implement a model architecture; perform initial training of the model; or train a custom model from an existing base model using additional data.
Service providers	Provide access to models for downstream applications together with the necessary computing power to run queries on the model.
Application developers	Develop user-facing applications that incorporate AI services.
Users	Interact with AI applications to generate information or content.
Other platforms	Disseminate and amplify generated content, e.g. social media platforms, online forums, or websites.

Any oversight framework should account for the variety of actors that contribute to an AI system. In some cases, a single entity may deploy compute resources, develop the model, and host a service or application. In other cases, the compute provider, model developer, and service or application providers may be different entities.

2. *Different models have different risk profiles*

Different models may have vastly different risk profiles. The risk of a model depends on a number of factors, including:

- **Size.** All else equal, a larger model (i.e. with many parameters) may be able to better understand complex relationships in data. As a result, larger models may deliver higher performance for specific tasks, and may be adaptable across a wider range of tasks.
- **Context.** A model with a larger "context window" can process larger inputs, such as long documents or many lines of code. These models may produce more complex outputs, such as more credible written or visual content, or more effective code.
- **Data quality.** All else equal, a model trained on limited or low quality data may be more prone to bias, errors, or omissions. These models may be affected by cultural, political, racial, or language bias; exhibit poor performance in domains not represented in the training data; or present misleading information as authoritative advice.
- **Compute resources.** A model with access to powerful computing may be able to support higher-capacity operations. These models may produce more content, more quickly, than a model running on regular hardware.
- **Customization.** A base model that is pre-trained on a public dataset may support a variety of tasks but exhibit average performance across these tasks. By comparison, custom models that are trained or "fine-tuned" with sensitive or proprietary data may exhibit higher performance for specific tasks.
- **Integration.** Models may be segregated from, or integrated with, other digital systems. An AI application that can access the Internet, publish to social media, or execute financial transactions may pose a higher risk than an application that runs the model in an isolated environment.

These factors must be evaluated together to assess the aggregate risk of an AI model, including the likelihood and consequences of misuse.

3. *There may be a range of mitigations for different risks*

We encourage policymakers to assess the full range of available mitigations before intervening with regulation. These may include a combination of consensus standards, novel features, licensing, or best practices. In addition, AI systems will be subject to existing technology-neutral rules governing liability, security, and privacy. Together, these mitigations can provide a layered defense to emerging risks.

For example, a range of mitigations can be implemented for safety. Safety encompasses many kinds of intentional and unintentional harm. AI could be intentionally misused for purposes such as abuse, fraud, plagiarism, or disinformation. AI could be exploited for purposes such as hyper-personalized advertising. Alternatively, AI may give rise to unintentional harm through inaccurate or biased outputs. In the first instance, these risks can be mitigated at a technical level.

Standards for content authenticity

For example, Stability AI is implementing content authenticity standards. Images generated through our hosted services can include metadata to indicate the content was produced with AI assistance.[13] Content authenticity standards can help users identify if they are interacting with AI content or AI bots, enabling them to exercise appropriate care and diligence.

Further, these standards can help social media platforms assess the provenance of content before amplifying it through their network. Platforms may develop more sophisticated risk-based criteria for upranking or downranking content using this metadata as a signal. This can help to prevent the viral spread of misinformation, and help to protect human creators from unfair mimicry or passing off by AI-generated content.

Proactive features to prevent harmful content

Other mitigations include the development of filters. For instance, in versions of Stable Diffusion developed exclusively by Stability AI, we filter training data to remove potentially unsafe images. By removing that data before it ever reaches the model, we can help to prevent users from generating harmful images in the first place. In addition, we filter content that is generated by users through our hosted applications or services. Combined with our ethical use license – which prohibits the unlawful or exploitative use of Stable Diffusion – these mitigations can help to prevent the misuse of AI for the production of harmful content.

Best practices for datasets

AI engages other issues beyond safety, such as fairness to creators whose content is included in training data. Training is an acceptable, transformative, and socially beneficial use of publicly-available content that is protected by fair use.[14] However, we support efforts to improve creator control over their content online.

For example, we are developing new ways to help creators qualify the use of their publicly-available content in AI training. For example, Stability AI has solicited opt-out requests from creators. Creators can indicate if they want to opt-out from AI training, and Stability AI has

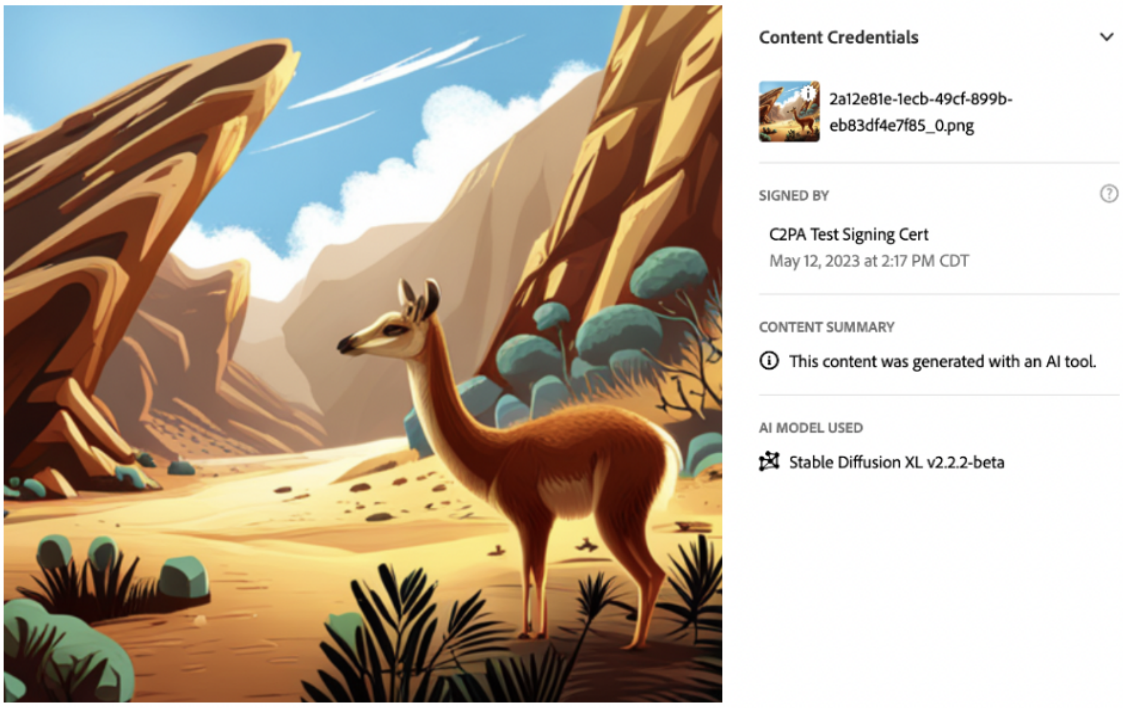
[13] Coalition for Content Provenance and Authenticity (C2PA) standards and the Content Authenticity Initiative (CAI), available here.

[14] Latent diffusion models are not a collage machine or search index for existing images. These models review pairs of images and text captions to learn the *relationship* between words and visual features, much like a student visiting an art gallery. With that acquired understanding, and creative direction from the user, these models can help users to generate new works.

committed to honoring these requests in the next wave of Stable Diffusion releases. Going forward, we are exploring technical standards for machine-readable opt-outs that follow the content wherever it goes. In addition, open datasets like LAION-5B respect digital protocols that indicate whether a website consents to automated data collection (e.g. robots.txt).

Improved techniques for training

In addition, we are implementing new techniques to improve diversity and reduce duplication in training data. These techniques can mitigate the risk of "overfitting", which occurs when a model erroneously and unintentionally overrepresents particular elements from an existing work (e.g. if an image model has only seen sunsets, it might learn that the sky is always orange). These measures can help to prevent users from replicating specific content with an AI model.



There are a range of mitigations for different risks. Above: An example of authenticity and provenance metadata indicating that an image was generated with an AI tool.[15]

[15] This example utilizes C2PA standards and an open-source interface from the CAI. See here.

Suggestions for the future of oversight

There is no silver bullet to address every risk in AI. Instead, we encourage policymakers to explore practical interventions that target specific, observable, emerging risks. Specifically, we encourage the Subcommittee to consider five suggestions for future consultation:

I. **Compute**. Larger models may pose a greater risk of misuse, adaptation, or weaponization by malicious state and nonstate actors. However, these models require significant compute resources for training and inference. Authorities may consider developing **disclosure or audit policies** that encourage cloud computing providers to report when their services are used for large scale or computationally intensive training and inference.[16] These reports can help to inform future policy.

II. **Models.** Stability AI is committed to the open development of transparent, accessible, and human-centric models. However, we recognize that extremely powerful models may require additional safeguards to prevent serious misuse. For example, authorities may consider **operational security and information security guidelines** for organizations that develop certain kinds of highly-capable and highly-adaptable models that meet predefined criteria for serious risk. In any case, authorities should continue to promote the public scrutiny of models through **open or independent evaluation**.

III. **Applications.** Users should know when they are interacting with AI. Authorities may consider **disclosure requirements** for application developers, and privacy obligations that require **affirmative user consent** prior to collecting their data for AI training. In addition, certain kinds of applications may pose a higher risk of harm to users, such as the provision of financial, medical, or legal advice, or the use of AI in administrative decisions. For these applications, regulators may consider robust **performance requirements** that describe evaluation criteria; required reliability; audit or assurance requirements; and interpretability requirements appropriate to the use-case.

IV. **Intermediaries**. Intermediaries such as social media platforms will continue to play a significant role in the dissemination of content online. These platforms can help to mitigate the risk of misinformation or disinformation from AI-generated content. Policymakers should promote the adoption of **content authenticity standards** by AI service and application providers, and encourage platforms to incorporate these signals as part of their content recommendation and moderation systems.

V. **Authorities.** The U.S. Government can intensify public investment in three areas. First, authorities should accelerate the development of **evaluation frameworks** for AI models in partnership with researchers, the developer community, and industry. Second,

[16] See, e.g., Article 56b in the European Parliament's draft AI Act (as adopted by committee vote on May 11, 2023) available here. Stability AI encourages policymakers to consult widely with the research community and industry on specific interventions.

policymakers should accelerate investment in **public compute and test bed resources**, implementing and expanding on the recommendations of the National AI Research Resource Taskforce.[17] These capabilities are essential to support the public research and public evaluation of AI. Third, policymakers should consider funding or procuring a **public foundation model**.[18] This model would be managed as a public resource; subject to public oversight; trained on trusted data; and available to organizations across the United States. A public model would support academic, small business, and public sector applications, helping to accelerate the safe adoption of AI across the economy.

These suggestions represent actionable, risk-based, and pro-innovation steps towards a future oversight framework. Where existing policy falls short, these measures may help to provide assurance to the public and guidance to industry.

Conclusion

The United States can demonstrate global leadership with a measured response to AI that realizes the full potential of these technologies while addressing emerging risks. Whatever the path forward, we encourage policymakers to vigorously promote openness in AI. These models will form the backbone of our digital economy, and it is essential that the public can scrutinize their development. Open models and open datasets will help to improve safety through transparency; foster competition in essential AI services; and ensure the United States retains strategic leadership in these critical technologies.

[17] NAIRR Task Force, 'Strengthening and Democratizing the U.S. AI Innovation Ecosystem', 2023.
[18] See, e.g., UK Government, 'Prime Minister and Technology Secretary announce funding for Foundation Model Taskforce', 2023 available here.