stability.ai

Statement to the U.S. Senate AI Insight Forum on Transparency, Explainability, and Copyright

Ben Brooks
Head of Public Policy
Stability AI

Thank you to Leader Schumer, Senator Rounds, Senator Heinrich, and Senator Young for the opportunity to contribute to this important dialogue. I am the Head of Public Policy for Stability AI, a leading developer of open technology helping to promote transparency and competition in AI. Our commitment to transparency extends to advocacy. We believe that difficult questions, from safety to copyright, are best scrutinized out in the open, and we welcome your leadership in shining a spotlight on these complex issues.

Background

Stability AI is a global company working to amplify human intelligence by making foundational AI technology accessible to all. Today, we develop AI models across a range of modalities, including image, language, audio, and video. Essentially, these models are software programs that can help a user to create, edit, or analyze complex content. With appropriate safeguards, we release these models openly, sharing our software code along with the billions of distinctive settings or "parameters" that define the model's performance. That means everyday developers and independent researchers can integrate or adapt our models to develop their own AI models, build their own AI tools, or start their own AI ventures, subject to our ethical use licenses.[1]

To date, our models have been downloaded over 100 million times by developers, and nearly 300,000 developers and creators actively contribute to the Stability AI online community.[2] Our family of image models, Stable Diffusion, underpin up to 80 percent of all AI-generated imagery.[3] These models can take a text instruction or "prompt" from a user and help to create a new image. In addition, we develop a suite of language models that can interpret, summarize, or generate text. These include highly capable large language models, compact language models, specialized models for software development, and models for underrepresented languages, including Japanese and Spanish. Our audio model, Stable Audio, generates high-quality soundtracks and was recently listed on the *TIME* Best Inventions of 2023. Building on this experience, we have developed video models that demonstrate new breakthroughs in video generation.[4] Further, we support academic research into scientific applications of AI. Stability AI provides a range of services to help partners customize and deploy our models, sustaining our open research efforts.

We are committed to the safe development of AI. To that end, we are signatories to the White House *Voluntary AI Commitments* and the British Government's *Joint Statement on Tackling Child Sexual Abuse in the Age of AI*; we participated in the first large scale public evaluation of AI models at DEF CON, facilitated by the White House, and the UK AI Safety Summit; and we engage with authorities around the world.

Open models promote transparency in AI

Generative AI will become critical infrastructure across the digital economy. These models will support creative, analytic, and scientific applications – from personalized tutoring to drug discovery – that go far beyond the caricature of "push a button, get an image" or "push a button, get a poem". Language models will power tools that revolutionize essential services, from education to healthcare; reshape how we search

---

[1] See e.g. the Open Responsible AI License (OpenRAIL) for Stable Diffusion, prohibiting a range of unlawful or misleading uses, available here. We use the term "open" to refer to any models with publicly-available parameters.
[2] Figures from Hugging Face and Discord, November 2023.
[3] Everypixel, 'AI Image Statistics', August 2023, available here.
[4] See e.g. Stability AI, 'Improving Latent Diffusion Models', July 2023, available here; Stability AI, 'Stable LM-3B Technical Report', October 2023, available here; Stability AI, 'Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets', November 2023, available here.

and access information online; and transform analysis, knowledge management, or decision making in some of our most important public and private sector institutions. Audiovisual models will power tools that radically accelerate the creative process, helping existing creators boost their productivity and experiment with new concepts while lowering barriers to entry for people who do not have the resources or training to realize their creative potential today. Instead of simply consuming the best available content, these "dormant" creators will be able to produce their best imaginable content.

It is more important than ever that we can scrutinize these systems before the next wave of digital services and digital ventures are built on "black box" technology operated by a small cluster of Big Tech firms. As Congress knows all too well, transparency in algorithmic technology is unfinished business. Today, our digital economy runs on opaque systems that feed us content on social media, control our access to information, determine our exposure to advertising, and mediate our online interactions. Everyday users and small businesses are unable to scrutinize these systems or build their own alternatives, and there is little competitive pressure on dominant firms to allow them to do so. Without a conscious effort to promote transparency and competition, AI is at risk of repeating that history. Against this backdrop, open models play a vital role in the emerging AI ecosystem:

- **Open models promote transparency.** Researchers and authorities can "look under the hood" of an open model to verify performance, identify risks or vulnerabilities, study interpretability techniques, develop new mitigations, and correct for bias. By comparison, closed models may not disclose how they are developed or how they operate. Closed models may be comparatively opaque, and risk management may depend on trust in the developer.

- **Open models lower barriers to entry.** Training a new "base" model from scratch requires significant resources that are not available to everyday developers. Open models lower these barriers to entry. Everyday developers can build on open models to create new AI tools or launch new AI ventures without spending tens of millions of dollars on research and computing.[5] In this way, the economic benefits of AI accrue to a broad community of developers and firms across the United States, not just Silicon Valley.

- **Open models drive innovation in safety.** Developers can refine open models for improved safety and performance in specific tasks. For example, open models can be optimized or "fine-tuned" through a range of techniques to mitigate undesirable behavior such as bias, misinformation, or toxicity. These techniques can yield significant improvements in the behavior of a model without requiring extensive computing resources. That means ordinary developers can build safer and more effective models to better support their real-world applications.

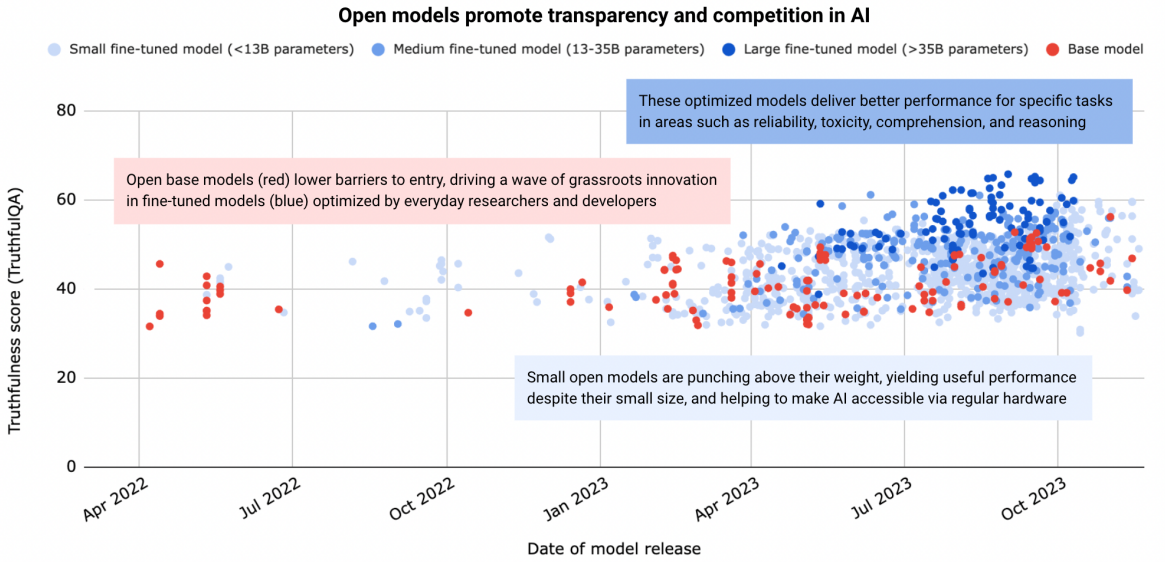
- **Open models foster strategic independence.** Open models enable public and private sector organizations to build independent AI capabilities without relying on a handful of firms for foundational technology. They can develop these AI capabilities securely "in house" without exposing their confidential data or ceding control of their distinctive model parameters to third parties. Operational independence will be important for organizations in sensitive or regulated sectors, such as healthcare, finance, law, and public administration.

- **Open models improve accessibility.** Many open models are smaller, more efficient, and more accessible than proprietary models. Unlike those models, which require significant computational resources to train and run, small open models can deliver useful performance with regular hardware. For example, open models may be hundreds of times smaller than a closed-source model such as GPT-4. Users can run small models on local devices, including smartphones, and developers can train or optimize these models with desktop hardware.

In this way, open models are fueling a wave of grassroots innovation in AI. Open models put this technology in the hands of everyday developers, independent researchers, and small businesses across

[5] OpenAI disclosed that it cost USD 100 million to train the closed-source GPT-4 model: Wired, 'Open AI's CEO says the age of giant models is already over', April 2023, available here.

America who are helping to build safer AI models and useful AI tools. Open models offer a transparent, competitive, and secure alternative to black box technology owned and operated by Big Tech firms.



Source: Our analysis of data from the Hugging Face "Open LLM Leaderboard" (November 2023). The TruthfulQA benchmark measures a model's tendency to reproduce falsehoods.

Transparency isn't a silver bullet, but there are layers of mitigation for risk

Transparency in open models helps to support AI safety. However, "AI safety" can mean many different things. In one sense, AI safety should be understood as a conventional product safety problem: whether a deployed AI system performs as expected or required for a given task. In this sense, open models support *risk mitigation* (i.e. the reduction of risk) by enabling developers and researchers to refine the behavior of a model before they deploy it in a user-facing system, such as a chatbot. For example, a raw or "pre-trained" base model might understand how to read, write, or draw, but it may be prone to undesirable behaviors such as bias, misinformation, or toxicity. It must be fine-tuned before deployment. Given access to the parameters of an open model, developers can adjust these behaviors before real-world deployment, taking into account their intended application and specific operating environment. In addition, open models support *risk assurance* (i.e. the verification of risks and mitigations) by enabling deployers, researchers, and authorities to directly scrutinize the behavior of a model. Where closed models depend on trust in the developer, open models can earn trust through transparency.

By itself, transparency is not a complete answer to risk mitigation and assurance. For example, interpretability remains a challenge – models can "reason" in unfamiliar or erroneous ways, and it can be difficult to understand how any model arrives at a particular output from a given input. In some cases, that can make it difficult to explain and justify the output to the user, which is a major limitation when AI is used for consequential decision making. Over time, we expect that performance-based evaluation through standardized testing will be essential to verify that a model operates as expected, and that it demonstrates the required level of reliability and robustness for a particular task. While evaluation is a developing field, further research into performance benchmarks, adversarial testing ("red-teaming"), and specialized human evaluation – including via the National Institute of Standards and Technology (NIST) and specific regulatory agencies – will help to provide confidence that AI systems deliver the expected or required performance.

We acknowledge that open models pose unique challenges for other kinds of AI safety, such as the prevention of misuse. For example, language models may be misused to generate intentional disinformation, exploit software vulnerabilities, or summarize dangerous information. Audiovisual models may be misused to generate misleading or unlawful deepfakes. As with other digital technologies, there are no silver bullets to eliminate the risk of misuse. However, there are layers of effective mitigations that help to make it easier to do the right thing with AI, and harder to do the wrong thing:

- As a first line of defense, **models** may be optimized for safer behavior prior to release through a range of techniques including data curation, instruction tuning, and reinforcement learning from human or AI feedback. For example, Stability AI filters unsafe content from our training data, helping to prevent the model from producing unsafe content. Further, we evaluate and fine-tune our models to help eliminate undesirable behaviors, such as sexualized or racialized bias. We disclose known risks and limitations in standardized formats to help downstream deployers decide on additional mitigations.

- As a second line of defense, **deployers** may filter unsafe prompts and unsafe outputs when they host a model through an application or interface. Stability AI implements these filters on our hosted services. In addition, we apply imperceptible watermarks and content provenance metadata to images generated through our interfaces. These signals can help social media platforms and search engines identify AI-generated content before amplifying it through their network. Eventually, these signals will help to inform more sophisticated content recommendation and content moderation systems. These efforts are detailed in our recent submission to the Federal Election Commission.[6]

- As a third line of defense, **users** are governed by technology-neutral rules that apply with equal force to the misuse of AI models (e.g. fraud, abuse, defamation, non-consensual intimate imagery, election interference, or hacking). Where necessary, these can be fortified to account for novel types of misuse or increased prevalence of misuse. For example, we have previously urged Congress to review legal guardrails governing the improper use of a person's physical or vocal likeness for misleading or exploitative purposes.

- As a fourth line of defense, AI **countermeasures** can be integrated across the digital economy to detect and defend against misuse. Today, AI models are used to detect unsafe content on social media and identify software vulnerabilities in complex security systems. Like conventional software, AI can be used as a shield, not just a sword, and we expect that defensive applications for AI will become increasingly effective in detecting, intercepting, and remediating various kinds of AI misuse.

No mitigation is watertight, but together, they provide a layered defense to emerging risks. As Congress considers the future of AI oversight, we encourage policymakers to adopt a holistic approach to serious misuse, consistent with the approach taken to other "dual use" technologies, from screwdrivers to software to satellite navigation systems. As with any technology, policymakers should (i) assess the incremental risk of catastrophic misuse, taking into account the realistic capabilities of AI models, (ii) measure the cumulative effectiveness of mitigations across the supply chain, both upstream and downstream, to determine the residual risk of catastrophic misuse, and (iii) weigh the benefits of open access against the opportunity costs of restrictive access. These are complex questions, but on the available evidence, we believe that existing AI should remain a presumptively open technology. Open innovation will best support the development of a transparent, competitive, and secure AI ecosystem.

Future policy should promote transparency in models

The best way to promote transparency in AI models is to promote diversity in the AI ecosystem. That ecosystem is more than a handful of corporate labs building closed products. It includes millions of developers, researchers, and creators who share and build on open technology. Their grassroots innovation is helping to make AI safe, useful, and accessible. However, prescriptive or overbroad statutory requirements for AI could have a chilling effect on that grassroots innovation, and we urge care in the development of rules directed at broad technology rather than specific harms. Instead, we encourage policymakers to:

[6] Stability AI, 'Comment on Petition for Rulemaking on AI in Campaign Ads', October 2023, available here.

- **Develop requirements proportional to risk.** The risk profile of an AI system depends on how the system is deployed. Policy should be risk-based, and account for these variations. For example, an AI system deployed in higher-stakes domains such as healthcare, finance, education, or public administration may attract more rigorous obligations than an AI system deployed in a lower-stakes domain such as entertainment (e.g. different requirements for reliability, interpretability, and robustness). In addition, policy should clearly distinguish between different kinds of risks, since these may have different mitigations. "One size fits all" requirements could hamper open innovation by imposing disproportionate or indiscriminate requirements on every AI system without accounting for the magnitude or type of risk.

- **Account for the variety of actors in the supply chain.** Models are just one component in an AI system. Different actors may train, fine-tune, host, deploy, and market different parts of a user-facing AI system. In that environment, policy should not assume vertical integration or formal relationships between actors in the supply chain, e.g. by imposing novel liability rules. Instead, liability should be determined through ordinary product liability principles, taking into account the distribution of responsibilities and the relationships between these actors. Further, policy should not assume that every actor is a sophisticated corporate entity. While firms may be able to comply with new statutory requirements, an everyday developer or independent researcher is unlikely to have the same resources or expertise. In particular, we urge caution in the development of mandatory disclosure, notification, audit, or licensing requirements. These measures would disproportionately burden developers and researchers who share or contribute to open models.

- **Invest in everyday risk, not just frontier risk.** Successfully integrating AI requires a sustained commitment to safety right across the AI ecosystem: from Big Tech to everyday developers, closed-source to open-source, and long-term threats to short-term risks. Research should not be limited to the most powerful models, and "trust and safety" shouldn't be the exclusive preserve of corporations. Governments can play a vital role in accelerating safety research across this diverse ecosystem. To that end, we welcome recent investments in research for conventional risks such as disinformation, privacy, and discrimination (through the new U.S. AI Safety Institute at NIST) in addition to research for catastrophic Chemical, Biological, Radiological, and Nuclear risks.

AI development furthers the objectives of copyright, but we acknowledge emerging concerns

We believe that AI development is an acceptable, transformative, and socially-beneficial use of existing content that is protected by fair use and furthers the objectives of copyright law, including to "promote the progress of science and useful arts".[7] Through training, generative AI models learn the unprotectable ideas, facts, and structures within a visual or language system, and that process does not interfere with the use and enjoyment of the original works. Free learning of these facts about our world is essential to recent developments in AI, and it is doubtful that these groundbreaking technologies would be possible without it. The United States has established global leadership in AI due, in part, to a robust, adaptable, and principles-based fair use doctrine that balances creative rights with open innovation.[8]

- **Models learn behaviors, they do not stitch together works.** During training, models learn the hidden relationships between words, ideas, and fundamental visual or textual features. The model doesn't rely on any single work in the training data, but instead learns by observing recurring patterns over vast datasets – much like a student visiting a library to learn how to read. These datasets consist of billions of image and caption pairs, trillions of words, or hundreds of years of video. A properly trained model does not store the works in this training data. They do not "collage" or "stitch" together existing works, nor operate as a "search engine" for existing content. The product of that training process is a piece of software that has learned certain behaviors and understands complex relationships.

[7] U.S. Constitution, Article I, Section 8, Clause 8.
[8] 17 U.S.C. §107.

- **Models apply this knowledge to new and unseen tasks.** Models can apply this learned knowledge to help develop new content or support new tasks that did not appear anywhere in the training data. This knowledge is generalizable, and models have a range of creative, analytic, and scientific applications that extend beyond simple generation: from editing photographs to identifying software bugs to developing new diagnostic approaches for complex medical disorders.

- **Models are components in a tool, not independent agents.** AI can help creators express themselves, but AI is not a substitute for creators. Instead, AI should be understood as a tool that can help to support the creative process. The AI model operates at the creative direction of the user, who provides instructions by supplying text prompts or reference examples, and by adjusting other settings. The user ultimately determines how the generated content is shared, displayed, and represented to others downstream.

Left: Generative AI models do not "stitch together" original works. They learn hidden relationships between words, ideas, and features within a visual, textual, or musical system. They apply this knowledge to help produce new works, and they often apply this knowledge imperfectly. For example, an AI-generated "Pentagon building" or "handshake" may appear to be highly realistic at first glance; on closer inspection, however, the AI-generated Pentagon has six or seven sides, not five, and the AI-generated hands may have two thumbs, or an irregular number of fingers.

We believe that existing legal frameworks effectively govern AI outputs, ranging from the replication of a specific work, to the use of protected likeness, to permissible experimentation with style. Likewise, existing frameworks can resolve questions of authorship. In principle, we acknowledge a threshold of authorship below which an AI-generated work with negligible human input may not qualify for registration. That said, we are concerned that recent U.S. Copyright Office (USCO) guidance and decisions may not account for the many ways in which human input can rise above that threshold. A user with clear expressive intent, who has demonstrated that they directed the AI system, should be able to register their work. We welcome further clarification on this issue. Overly discretionary guidance means that creators may be unfairly disadvantaged by their use of AI tools within a wider creative workflow.

We recognize the concern among some creators about the development and deployment of these systems. We are actively working to address these concerns through technology, standards, and good practices. These efforts, including opt-outs, labeling, training, and data access, are detailed in our recent submission to the USCO.[9] In general, as we integrate AI tools into the digital economy, we believe the community will continue to value human-generated works – and perhaps value them at a premium. Smartphones didn't destroy photography, and word processors didn't diminish literature, despite radically transforming the economics of creation. Instead, they gave rise to new demand for services, new markets for content, and new creators. We expect the same of AI tools, and we welcome an ongoing dialogue with the creative community about the fair integration of these technologies.

Conclusion

AI models will be the backbone of our digital economy, and it is essential that the public can scrutinize their development. As part of the diverse AI ecosystem, open models will advance safety through transparency, foster competition, and ensure the United States retains strategic leadership in AI adoption. Grassroots innovation is America's greatest asset, and open models put these technologies in the hands of everyday developers, independent researchers, and small businesses who can help turn AI into useful tools that amplify human intelligence.

[9] Stability AI, 'Response to the Inquiry into AI and Copyright', October 2023, available here. See also our testimony to the Senate Judiciary Subcommittee on Intellectual Property, July 2023, available here.