

Artificial Intelligence and Intellectual Property – Copyright

Senate Judiciary Subcommittee on Intellectual Property

Wednesday 12 July

Ben Brooks

Head of Public Policy

Stability AI

<i>Introduction</i>	1
<i>Background</i>	2
<i>Stability AI develops open technology to drive transparency, competition, and innovation in AI</i>	2
<i>We believe AI is a tool, not a substitute for creators</i>	4
<i>AI models learn the relationships between words, concepts, and visual or textual features</i>	5
<i>How AI is trained, deployed, and used has implications for future policy</i>	7
<i>We are actively working to address emerging concerns and mitigate risks</i>	9
<i>Conclusion</i>	10

Introduction

Thank you, Chair Coons and Ranking Member Tillis. We welcome the opportunity to appear today, and to share our perspective on these important issues. AI is a tool that can help to unlock creativity, drive innovation, and open up new opportunities for creators and developers across the United States. With appropriate safeguards, we are working to make these foundational technologies accessible to all. As with any groundbreaking technology, AI raises important questions about the integration of these tools into our everyday lives, including in the creative economy. We recognize these challenges and the depth of concern among creators. While we don't have all the answers, we are committed to an open and constructive dialogue, and we are actively working to address these concerns through technology, standards, and best practices.

Background

Stability AI is a global company that aims to unlock humanity’s potential by making foundational AI technology accessible to all. Today, Stability AI develops a variety of generative AI models. These models are software programs that analyze vast datasets to learn the relationships between words, ideas, and fundamental visual or textual features. They can apply this knowledge to help a user generate new content. In 2022, we took over the exclusive development of Stable Diffusion, a popular image model that takes a text instruction or “prompt” from a user and helps to produce new images. In 2023, we released the first in a series of language models known as StableLM to support open research into AI safety. These language models can take a prompt from a user and help to produce new passages of text or software code.

By some measures, developer interest in Stable Diffusion has grown faster than many open-source software projects in recent history, and over 200,000 creators and developers actively contribute to the Stability AI community. Stability AI partners with organizations to adapt these models for specific purposes, helping to sustain our open research and development efforts. In addition, Stability AI provides services to help developers and users access the computing resources necessary to train or run these models, and actively supports research into scientific applications of AI.

Stability AI develops open technology to drive transparency, competition, and innovation in AI

With appropriate safeguards, we release our flagship AI models openly as a public resource, subject to ethical use licenses. That means we publish the software code along with the billions of distinctive settings (“parameters”) that define the model’s performance. Creators and developers can freely adapt these open models. For example, they can customize their own AI models for specialized tasks, power their own AI applications, or build their own AI businesses. Open models are different to closed models, which are software programs that cannot be inspected or adapted by third-parties. Closed models may not disclose how they are trained or how they operate.

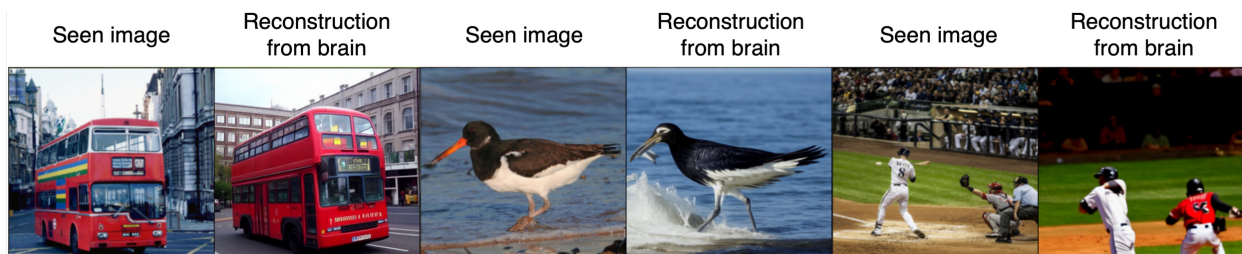
Open models can help to promote transparency, competition, and innovation in the digital economy. Within the AI ecosystem, open models play an important role in several ways:

1. **Promoting transparency.** AI models will form the backbone of our digital economy, and everyone should have a voice in their design. They will power chatbots that provide medical advice or personalized tutoring; transform how we search, access, or manage information; and support high-stakes analysis or decision-making in some of our most important public and private institutions. By releasing open models, researchers, authorities, and developers can “look under the hood” to verify the performance and suitability of a model for a particular task. They can identify potential risks, and help develop new mitigations. Likewise, they can inspect open datasets for quality, fairness, or

bias. Because these models are transparent, firms and agencies in sensitive sectors can build on open models to produce their own specialized models for particular tasks.¹

- 2. Fostering competition.** Developing a generative AI model typically requires significant resources. Open models can lower these barriers to entry, fueling competition in AI.² Using open models, creators and developers can deploy new AI tools or launch new AI ventures without spending millions on research and computing power. They can participate in this new industrial revolution as builders – not just consumers – of AI technology, and they can do so without relying on a handful of firms for critical infrastructure.
- 3. Supporting grassroots innovation.** Grassroots innovation by anyone, anywhere is one of America’s greatest assets, and open models put these capabilities in the hands of everyday creators, developers, and researchers. Today, models like Stable Diffusion are used for everything from editing photographs to prototyping architectural designs to researching new techniques for diagnosing complex disorders. We can’t anticipate every possible use of AI, but everyday people can experiment with open models to develop new and innovative applications that support their work and serve their community. In this way, open models can help distribute the economic benefits of AI across the United States, beyond Silicon Valley.

We build models to support and augment our users, not replace them. We develop tools that help everyday people and everyday firms. We are focused on practical AI capabilities that can be applied to everyday tasks – not a quest for an alien superintelligence. Designing around these principles can help to unlock the useful potential of AI while minimizing the risk of misuse, weaponization, or “runaway” systems.



Above: Multiple teams are experimenting with image models like Stable Diffusion to research new techniques for studying medical disorders. Here, MedARC researchers demonstrate how latent diffusion models can be used to visualize how patients are perceiving a given scene, based only on fMRI scans. Disturbances in perception could help to assess or diagnose complex neurological conditions.³

¹ For example, a regulated financial institution may customize AI models to assist in analysis, decision making, or customer support. The financial institution may need 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 the AI model without relying on a third-party provider. By building on open models, a financial institution can train and manage their own AI system.

² See, e.g. the Hugging Face “leaderboard” comparing open language models, available [here](#).

³ MedARC projects are available [here](#).

AI is a tool, not a substitute for creators

We believe that AI is a tool that can help creators express themselves, but it is not a substitute for creators. Instead, AI can help to accelerate the creative process. AI tools can help existing creators boost their productivity, experiment with new concepts, and perform complex tasks as part of a wider workflow. In addition, AI can lower barriers to entry for people who do not have the resources or training to realize their creative potential. Instead of simply consuming the best available content, these “untapped” creators will be able to produce their best imaginable content. As with other assistive technologies – from paintbrushes to cameras to editing software – the creator retains ultimate control over the composition and use of their work.

In their own words

"As a designer, I've used Stable Diffusion to visualize ideas in a matter of minutes. I'll use it as a tool to create different iterations of an idea and manipulate quickly. I'll create moodboards and storyboards with it for other team members to see where we are headed."

– Stefania Bulbarella, Broadway Projection Designer

"As a designer and educator working with generative AI, I have consistently observed that groundbreaking innovations primarily originate from open-source platforms. The ability to fine-tune models and weights through the collaborative efforts of diverse user communities is a critical counterpoint to the potential biases and decisions imposed by closed-source AI corporations. Therefore, advocating for the widespread adoption and endorsement of open-source AI frameworks is imperative, as it not only fortifies the United States' standing as the forefront of AI research and development but also stimulates the growth, resilience, and variety within the field."

– Andrew Kudless, Architect

"In the past year, I've used Stable Diffusion as an early conceptual design exploration tool in my contemplative architectural designs and also as a moodboard dynamic library of original reference images to visually communicate lighting design ideas in similar spaces with similar materials."

– Ilva Dodaj, Architect and Lighting Designer

"The tool fits into my workflow such that I can spend more time doing what I love most, which is designing thoughtful solutions for spaces that will empower its users. Generative AI does not threaten the taste, reasoning, and empathy required to do that."

– Isabella Orsi, Interior Designer⁴

⁴ Feedback from Bulbarella, Kudless, and Dodaj was kindly shared with Stability AI over the past few days. Orsi shared feedback on Twitter and with the [New York Times](#).

AI models learn the relationships between words, concepts, and visual or textual features

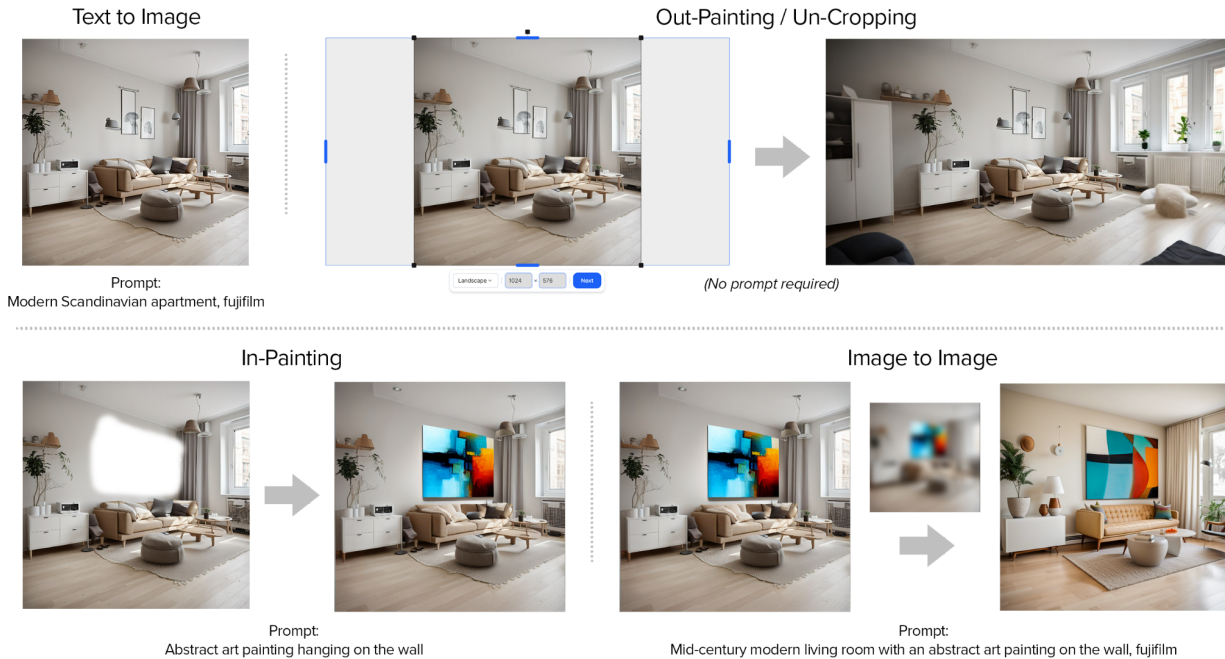
Recent AI models are described as “generative” AI because they can generate complex new content, helping to simplify analytic or creative tasks. These models analyze vast datasets to understand the relationships between words, concepts, and visual or textual features – much like a student visiting a library or an art gallery. Models can then apply this knowledge to help a user produce new content. This learning process is known as training.

Image

For example, during training, an image model such as Stable Diffusion will review billions of pairs of images with associated text captions.⁵ Through this process, the model learns to identify fundamental visual structures within images, such as shapes, textures, and patterns. By cross-referencing with known text captions, the model learns to associate these fundamental structures with particular terms. For example, the model may learn to understand the appearance of fur on a “dog”; learn how light interacts with “water”; or capture the visual aesthetic described by words like “bleak” or “Renaissance”. When a user inputs a prompt – such as “a photorealistic astronaut riding a pig” – the model can help to express the desired features in a new image, even if the model has never seen an example of that composition.

Image models are typically trained on large datasets. For example, Stable Diffusion was trained on a filtered subset of two billion image and caption pairs from LAION-5B, an open dataset of publicly-accessible image links and captions compiled by the European nonprofit Large AI Open Network (LAION). Following initial training, the model can be further trained or customized with additional data to improve performance in specific tasks.

⁵ Stable Diffusion is a type of latent diffusion model.



Above: Image models like Stable Diffusion can be used in a range of ways as part of a design workflow. They can help to produce new images based on a text description, fill in or replace parts of an existing image, extend parts of an existing image, or transform an existing image.

Language

Likewise, recent language models are trained on vast quantities of existing text. The model is typically shown incomplete fragments of text, broken into digestible units known as tokens, and it then studies the relationships between and among the tokens to detect complex patterns in how sentences, paragraphs, and passages are structured in different contexts.

In this way, the model learns underlying structures within a language system. For example, the model can identify the typical arrangement of a “blog”, “email”, or “speech”. It can understand the expected structure of an “example” or an “explanation”, and the textual features that contribute to tones like “funny” or “formal”. The model can learn what it means to distinguish, summarize, expand, or combine different concepts, such as “banana”, “Wall Street”, or programming instructions like “print()”. When a user inputs a prompt – such as “suggest some ways to improve this draft greeting card” – the model can interpret the request, and respond with a passage of text.

Language models are typically trained on datasets consisting of a trillion or more tokens of content (equivalent to hundreds of billions of words). Following initial training, these models can be “fine-tuned” to improve performance in specific domains, and to mitigate potential risks such as bias, misinformation, or toxicity. For example, reinforcement learning from human feedback

(RLHF) can help to improve the performance of a model by using human feedback to “rate” the model’s outputs. Likewise, instruction-tuning techniques can help to improve performance by providing labeled examples of sample inputs and desired outputs. Stability AI is actively engaged in research to improve safety through fine-tuning. This work draws on a number of open datasets shared by researchers for this purpose.

U I'm not a native English speaker. I find it hard to understand grammar. Can you help me by explaining prepositions?

S Sure, I'd be happy to help you understand prepositions!

Prepositions are words that show the relationship between a noun or pronoun and other words in a sentence. They are used to indicate location, direction, time, manner, and other relationships.

Here are some common prepositions and their meanings:

- "in": indicates location within an area, such as "I am sitting in the living room."
- "on": indicates location on an object, such as "I am holding the book on my lap."
- "at": indicates location at a specific time or place, such as "I met him at the coffee shop."
- "by": indicates a means of transportation or a means of achieving something, such as "I got there by walking."
- "with": indicates association or accompaniment, such as "I am friends with John."
- "from": indicates origin, such as "I am from the United States."
- "to": indicates direction, such as "I am going to the store."

Prepositions are usually used before a noun or pronoun, but they can also be used in other positions in a sentence, such as at the beginning or end.

U Can you help me implement a bubble sort in Python?

S Sure, here's an example implementation of bubble sort in Python:

```
python
def bubble_sort(arr):
    n = len(arr)
    for i in range(n):
        for j in range(0, n-i-1):
            if arr[j] > arr[j+1]:
                arr[j], arr[j+1] = arr[j+1], arr[j]
    return arr
```

This function takes in an array `arr` as input and returns the sorted array. The algorithm works by repeatedly swapping adjacent elements if they are in the wrong order. The `n` variable is set to the length of the array, and the outer loop runs `n` times, while the inner loop runs `n-i-1` times for each iteration of the outer loop, where `i` is the current iteration number. This ensures that the largest element is compared and swapped with the first element of the inner loop, and so on, until the array is sorted.

Above: Language models can be used in a range of ways as part of a creative workflow. For example, they can help to draft or edit documents or correspondence, or help to identify bugs in software code and suggest improvements.

How AI is trained, deployed, and used has implications for future policy

The training, deployment, and use of AI models has implications for how we should think about the interaction of AI tools with copyright. Future policy should account for several characteristics:

1. **Training data is not stored in the model.** Through training, these models develop an understanding of the relationship between words, concepts, 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 (billions of image and caption pairs, and hundreds of billions or trillions of words). The model does not store the material in this training data. They do not “collage” or “stitch” together original works, nor do they operate as a “search engine” for existing content.
2. **These models help users to produce new works.** Models apply this knowledge to help a user generate new and unseen content. That could mean a novel image, passage of text, block of code, series of instructions, or video clip.⁶ This knowledge is generalizable, which means it can help to develop new content and support new tasks that did not appear in the training data.

⁶ In addition, these models can be used as analytic tools to detect and classify features within content, such as visual patterns, textual information, or software bugs.

3. **The model is a creative tool, not an independent agent.** The model is a tool that helps to produce this content at the request of a user. The user provides creative direction by supplying text prompts or reference examples, and adjusting other settings. The user ultimately determines how the generated content is shared, displayed, or represented to others downstream.
4. **Models may be developed, customized, and deployed by different entities.** Models may be released as free and open software to promote transparency, competition, and innovation. The developer of a free and open model may have no relationship with downstream developers. For example, a creator, researcher, or business can choose to fine-tune these models for specialized tasks by training on their own data. In this environment, the original model developer may have no financial or contractual relationship with those who actually customize or deploy the model.

Training these models is an acceptable, transformative, and socially-beneficial use of existing content that is protected by the fair use doctrine and furthers the objectives of copyright law, including to “promote the progress of science and useful arts”.⁷ These models learn the unprotectable ideas, facts, and structures within a visual or textual system, and that process does not interfere with the use and enjoyment of 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 US 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. Other jurisdictions, including Singapore, Japan, and the European Union, have begun to incrementally revise their copyright laws to create safe harbors for AI training that achieve similar effects to fair use.



Above: These models do not “stitch together” original works. For example, image models notoriously struggle with hands. In this case, the model produces a nine-fingered hand because it hasn’t understood that a hand typically has five fingers, and it isn’t searching a database of images with hands. Instead, it has learned that a “hand” means a flesh-colored artefact associated with some number of appendages.

⁷ U.S. Constitution, Article I, Section 8, Clause 8.

We are actively working to address emerging concerns and mitigate risks

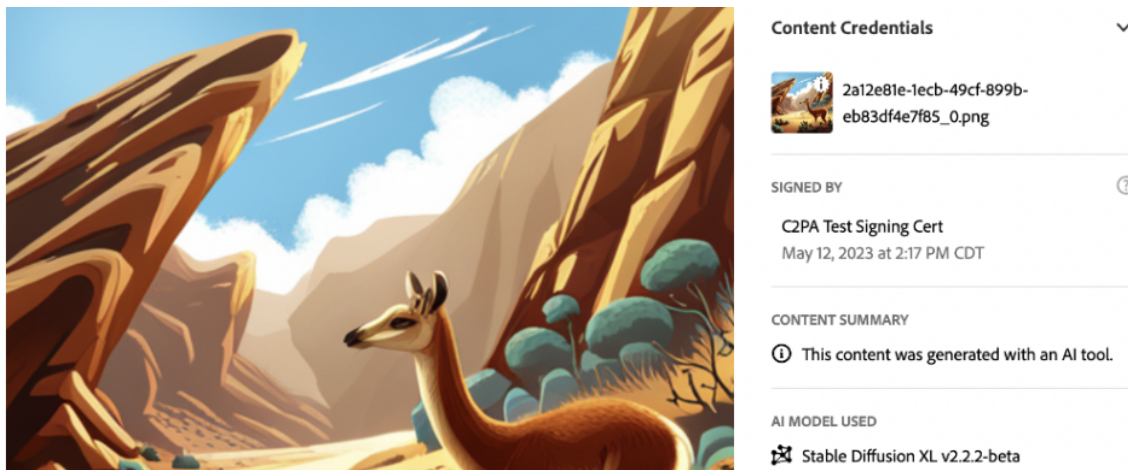
We understand that developments in AI can feel highly disruptive, and raise important questions about the application of existing law and public expectations to new technology. We are actively working to address emerging concerns. In particular, we support greater transparency in how AI content is used, produced, and shared through the digital economy. To that end, we are focusing our efforts on a number of initiatives:

1. **Best practices in training.** Because Stable Diffusion is pre-trained with content from open datasets, creators can determine whether their works appear in those datasets.⁸ Stability AI has proactively solicited opt-out requests from creators, and will honor these over 160 million opt-out requests in upcoming training for new Stable Diffusion models. Going forward, we are exploring new technical standards for machine-readable opt-outs, so that opt-out metadata follows the content wherever it goes. In addition, the datasets used for training Stable Diffusion respected industry-standard digital protocols like robots.txt, which indicates whether a website consents to automated data collection for ancillary purposes such as indexing or analysis.
2. **Identifying AI content.** We are implementing content authenticity standards and watermarking that can help users and content platforms better identify AI-assisted content.⁹ Images generated through our hosted API can include metadata and imperceptible watermarks to indicate the content was produced with an AI tool. The metadata is certified with the relevant model and version number. Downstream intermediaries – such as social media platforms – can use metadata, watermarks, and other signals to assess the provenance of content before amplifying it through a network. Together, these features can help platforms to distinguish AI content, enable users to exercise appropriate care when interacting with AI content, and help to limit the spread of misinformation with AI tools.
3. **Quality of datasets.** We are working to improve the quality of datasets to avoid unintended behavior in our models. For example, improving diversity and reducing duplication in training data can help to prevent “overfitting”, a training defect which occurs when a model erroneously overrepresents elements of a particular image from the dataset (e.g. if a model has only seen sunsets, it might learn that the sky is always orange). These measures can help to mitigate the risk of a user prompting a model with sufficient detail or determination to generate a lookalike work. Additionally, datasets that are more representative of diverse culture, language, demographics, and values can help to mitigate the risk of bias in outputs. We are actively investing in efforts to improve representation in these datasets.

⁸ See, e.g. the “Have I Been Trained?” portal via Spawning.AI, available [here](#).

⁹ In partnership with the Content Authenticity Initiative led by Adobe, available [here](#), we are implementing the C2PA standard for content metadata.

4. **Mitigations for misuse.** We have implemented a range of mitigations to prevent other forms of misuse. For example, we impose licensing conditions that balance open innovation with harm minimization. Our flagship image model, Stable Diffusion, is governed by the Open Responsible AI License (Open RAIL), which prohibits a range of unlawful or unethical uses.¹⁰ We publish standardized information about the performance and limitations of our models (see **Annex A**). In versions of Stable Diffusion developed exclusively by Stability AI, we apply filters on training data to remove unsafe images. By removing that data before it ever reaches the model, we can help to prevent users from generating harmful images. In addition, on Stability AI applications and APIs, we have developed and implemented effective filters that block unsafe text prompts or generated images that violate our platform terms of service.¹¹ In our language model research, we are exploring techniques to improve the screening of training data for personally identifiable information to mitigate the risk of inadvertent re-disclosure.



Above: An example of content authenticity metadata indicating an image was generated with an AI tool.

Conclusion

AI is a tool that will unlock creativity, innovation, and productivity. With appropriate safeguards, open models can help to put these tools in the hands of everyday creators and developers. As AI is integrated into the digital economy, we believe the community will continue to value human-generated content. Indeed, that content may be valued 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 will be true of AI tools, and we welcome an ongoing dialogue with the creative community about the fair deployment of these technologies.

¹⁰ Open Responsible AI License (OpenRAIL) available [here](#). Our license prohibits the unlawful, exploitative, or misleading use of Stable Diffusion.

¹¹ See, e.g. terms of service for the API available [here](#).

Annex A. Stability AI releases standardized information describing the performance, limitations, and biases of our models, in addition to other disclosures such as environmental impact. Examples of these disclosures include:¹²

<p>Misuse and Malicious Use</p> <p>Using the model to generate content that is cruel to individuals is a misuse of this model. This includes, but is not limited to:</p> <ul style="list-style-type: none"> • Generating demeaning, dehumanizing, or otherwise harmful representations of people or their environments, cultures, religions, etc. • Intentionally promoting or propagating discriminatory content or harmful stereotypes. • Impersonating individuals without their consent. • Sexual content without consent of the people who might see it. • Mis- and disinformation • Representations of egregious violence and gore • Sharing of copyrighted or licensed material in violation of its terms of use. • Sharing content that is an alteration of copyrighted or licensed material in violation of its terms of use. 	<p>Limitations</p> <ul style="list-style-type: none"> • The model does not achieve perfect photorealism • The model cannot render legible text • The model does not perform well on more difficult tasks which involve compositionality, such as rendering an image corresponding to “A red cube on top of a blue sphere” • Faces and people in general may not be generated properly. • The model was trained mainly with English captions and will not work as well in other languages. • The autoencoding part of the model is lossy • The model was trained on a subset of the large-scale dataset LAION-5B, which contains adult, violent and sexual content. To partially mitigate this, we have filtered the dataset using LAION’s NFSW detector (see Training section).
<p>Bias</p> <p>While the capabilities of image generation models are impressive, they can also reinforce or exacerbate social biases. Stable Diffusion was primarily trained on subsets of LAION-2B(en), which consists of images that are limited to English descriptions. Texts and images from communities and cultures that use other languages are likely to be insufficiently accounted for. This affects the overall output of the model, as white and western cultures are often set as the default. Further, the ability of the model to generate content with non-English prompts is significantly worse than with English-language prompts. Stable Diffusion v2 mirrors and exacerbates biases to such a degree that viewer discretion must be advised irrespective of the input or its intent.</p>	<p>Environmental Impact</p> <p>Stable Diffusion v1 Estimated Emissions Based on that information, we estimate the following CO2 emissions using the Machine Learning Impact calculator presented in Lacoste et al. (2019). The hardware, runtime, cloud provider, and compute region were utilized to estimate the carbon impact.</p> <ul style="list-style-type: none"> • Hardware Type: A100 PCIe 40GB • Hours used: 200000 • Cloud Provider: AWS • Compute Region: US-east • Carbon Emitted (Power consumption x Time x Carbon produced based on location of power grid): 15000 kg CO2 eq.

¹² See ‘Stable Diffusion V2-1 Model Card’ available [here](#).