

Common Sense Media AI Risk Assessment:

Stable Diffusion

Powerful image generator can unleash creativity, but is wildly unsafe and perpetuates harm

Last updated: Aug 6, 2024

Overall risk level: High ▾

Type of AI: Multi-Use ▾

Type of Review: Product Review ▾

Other Information:

- [How We Review](#)

⚠️ **DISCLAIMER:** We will not link directly to Stable Diffusion, DreamStudio, or Stability AI in this review, as we do not consider this a safe tool in any way. [Why this matters](#).

Key Takeaways

1. **Stable Diffusion is an generative AI product created by Stability AI that can create realistic images and art from text based descriptions.** Stable Diffusion can make authentic modifications to user-uploaded images, incorporate or eliminate components such as shadows or textures, and extend images beyond the original frame. It has the potential to enable creativity and artistic expression, allow for visualization of new ideas, and create new concepts or campaigns, yet serious risks remain.
2. **Stable Diffusion is susceptible to generating images that perpetuate harmful stereotypes, shaping impressionable minds with little accountability.** Even when directed to avoid bias, it can produce images that reinforce harmful stereotypes regarding race and gender—an issue we repeatedly confirmed through testing. Repeated exposure to biased worldviews can distort perceptions of people and society, particularly for kids and teens, and can influence how children understand what is “correct” and shaping their real-world interactions.

3. **Stable Diffusion has been used to create child sexual abuse material (CSAM).**
Stable Diffusion has already been used to create lifelike images of child sexual abuse, including sexual abuse of babies and toddlers, which have been sold online. While Stable Diffusion’s most recent update has aimed at preventing the generation of CSAM, the open source nature of the model—the fact that the model is freely available for anyone to use or modify—allows for easy removal of those protections.
4. **Inappropriate sexualized representations of women and girls harms all users.**
Stable Diffusion continues to produce harmful stereotypes, unfair bias, unrealistic ideals of women’s beauty and “sexiness”, and incorrect beliefs around intimacy for humans of all genders. Even with deliberate testing seeking images of women professionals, it easily produces inappropriately sexualized representations of women and girls. Research shows that the greater exposure to images that promote the objectification of women adversely affects the mental and physical health of girls and women.
5. **Stable Diffusion can be used to spread misinformation, exploit people’s likeness, and harm individuals.** While innovative and useful in many contexts, the high degree of freedom to alter images means Stable Diffusion’s features present new risks. Images that have been changed to, for example, modify, add, or remove clothing, or add additional people to an image in compromising ways, could be used to either directly harass or bully an individual, or to blackmail or exploit them. These features can also be used to spread disinformation campaigns and intentionally misinform others.

Table of Contents

Common Sense Media AI Risk Assessment.....	1
Key Takeaways.....	1
Table of Contents.....	2
Common Sense Media Risk Assessment.....	3
What is Meta AI?.....	3
What every parent needs to know.....	4
How to protect your teen right now.....	4
The biggest risks.....	5
Where it's best.....	14
Common Sense Media AI Principles Assessment.....	15
Keep Kids & Teens Safe: Unacceptable Risk.....	15

<u>Be Effective: Unacceptable Risk.....</u>	<u>17</u>
<u>Prioritize Fairness: High Risk.....</u>	<u>18</u>
<u>Put People First: Unacceptable Risk.....</u>	<u>21</u>
<u>Support Human Connection: Unacceptable Risk.....</u>	<u>21</u>
<u>Be Trustworthy: High Risk.....</u>	<u>22</u>
<u>Use Data Responsibly: Moderate Risk.....</u>	<u>23</u>
<u>Be Transparent & Accountable: High Risk.....</u>	<u>24</u>
<u>References.....</u>	<u>25</u>

Common Sense Media Risk Assessment

What is Stable Diffusion?

Stable Diffusion is a generative AI product created by Stability AI. It can create realistic images and art from a text-based description that can combine concepts, attributes, and styles. Stability AI's full suite of image editing tools offers users a sophisticated range of options: extending generated images beyond the original frame (outpainting), making authentic modifications to existing user-uploaded or AI-generated pictures, and incorporating or eliminating components while considering shadows, reflections, and textures (inpainting). Once users achieve the generated image they want, they can download and use it.

Stability AI released Stable Diffusion to the public in November 2022. It is powered by a massive data set of image-text pairs scraped from the internet. The data set includes a subset of 2.32 billion images that contain English text. It was created by LAION, which stands for "Large-scale Artificial Intelligence Open Network." LAION is a nonprofit organization that is funded in part by Stability AI.

Stability AI's hosted version of Stable Diffusion can be accessed via its cloud service DreamStudio. DreamStudio extends beyond text-to-image prompting by providing inpainting, outpainting, and image-to-image generation. Users purchase credits to pay for the computing cost of each request. Currently, \$10 equals 1,000 credits, which Stability AI notes is ~5,000 images.

In addition, Stable Diffusion has made all model weights and code available. Anyone is able to access, download, and use the full model.

How it works

Stable Diffusion is a form of generative AI, which is an emerging field of artificial intelligence. Generative AI is defined by the ability of an AI system to create ("generate") content that is complex *and* coherent *and* original. For example, a generative AI model can create sophisticated writing or images.

Stable Diffusion uses a particular type of generative AI called "diffusion models," named for the process of diffusion to generate new content. Diffusion is a natural phenomenon you've likely experienced before. A good example of diffusion happens if you drop some food coloring into a glass of water. No matter where that food coloring starts, eventually it will spread throughout the entire glass and color the water in a uniform way. In the case of computer pixels, random motion of those pixels will always lead to "TV static." That is the

image equivalent of food coloring creating a uniform color in a glass of water. A machine-learning diffusion model works by, oddly enough, destroying its training data by successively adding "TV static," and then reversing this to generate something new. They are capable of generating high-quality images with fine details and realistic textures.

Stable Diffusion combines a diffusion model with a text-to-image model. A text-to-image model is a machine learning algorithm that uses natural language processing (NLP), a field of AI that allows computers to understand and process human language. Stable Diffusion takes in a natural language input and produces an image that attempts to match the description.

Where it's best

- Stable Diffusion has the potential to enable creativity and artistic expression, allow for visualization of new ideas, and create new concepts and campaigns.
- Stability AI suggests that the best uses of Stable Diffusion include: generation of artworks and use in design and other artistic processes; applications in educational or creative tools; research on generative models; safe deployment of models that have the potential to generate harmful content; and probing and understanding the limitations and biases of generative models.

The biggest risks

- **Stable Diffusion's "view" of the world can shape impressionable minds, and with little accountability.** Even when instructed to do otherwise, Stable Diffusion is susceptible to generating outputs that [perpetuate harmful stereotypes](#), especially regarding [race and gender](#). We confirmed this repeatedly with our own testing. These propensities toward harm are frighteningly powerful. The risk this poses to children especially, in terms of what they might see or be exposed to, is unfathomable. What happens to our children when they are exposed to the worldview of a biased algorithm repeatedly and over time? What view of the world will they assume is "correct," and how will this inform their interactions with real people and society? Who is accountable for allowing this to happen?
- **Stable Diffusion has been used to create child sexual abuse material (CSAM).** Stable Diffusion has been used to create lifelike images—[sometimes many thousands of them by a single bad actor](#)—of child sexual abuse, including of the sexual abuse of babies and toddlers. [These images have then been sold online](#). While Stable Diffusion's July 2023 update aimed to prevent it from generating some of the most objectionable content, the open source nature of the model

allows for easy removal of those protections, or for older versions to be used, in applications built from the technology.

- **Inappropriate sexualized representations of women and girls harm all users.** Despite many public failings, Stable Diffusion continues to easily produce inappropriately sexualized representations of women and girls, even with prompts seeking images of women professionals. This perpetuates harmful stereotypes, unfair bias, unrealistic ideals of women's beauty and "sexiness," and incorrect beliefs around intimacy for humans of all genders. Numerous studies have shown that greater exposure to images that promote the objectification of women [adversely affects the mental and physical health](#) of girls and women. Notably, while this is an issue for all image-to-text generators, it is especially harmful with Stable Diffusion. This is because of the combination of an uncensored data set and minimal protections, such as a refusal to generate images when it detects prompts that violate the company's terms of service.
- **Stable Diffusion consistently and easily reinforces harmful stereotypes.** While Stable Diffusion's July 2023 update aimed to prevent it from generating some of the most objectionable content, this remains a significant risk. Recent findings show continued reinforcement of harmful stereotypes, and the manner in which Stability AI has open-sourced the model allows anyone to remove those protections in new applications. A great resource for exploring this problem further can be found at [Stable Bias](#). Relevant articles:
 - Tikun, N., Schaul, K., & Chen, S.Y. (2023, Nov. 1). [How AI is crafting a world where our worst stereotypes are realized](#). *Washington Post*.
 - Crawford, A., & Smith, T. (2023, June 28). [Illegal trade in AI child sex abuse images exposed](#). BBC.
 - Harlan, E., & Brunner, K. (2023, June 7). [We are all raw material for AI](#). BR24.
 - Nicoletti, L., & Bass, D. (June 2023). [Humans are biased. Generative AI is even worse](#). *Bloomberg*.
 - Vincent, J. (2023, Jan. 16). [AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit](#). *The Verge*.
 - Edwards, B. (2022, Sept. 21). [Artist finds private medical record photos in popular AI training data set](#). *Ars Technica*.
 - Wiggers, K. (2022, Aug. 24). [Deepfakes for all: Uncensored AI art model prompts ethics questions](#). *TechCrunch*.
 - Wiggers, K. (2022, Aug. 12). [This startup is setting a DALL-E 2-like AI free, consequences be damned](#). *TechCrunch*.
- **Stable Diffusion's advanced inpainting and outpainting features present new risks.** While innovative and useful in many contexts, the high degree of freedom to

alter images means they can be used to perpetuate harms and falsehoods. Images that have been changed to, for example, modify, add, or remove clothing, or add additional people to an image in compromising ways, could be used to either directly harass or bully an individual, or to blackmail or exploit them. These features can also be used to create images that intentionally mislead and misinform others. For example, disinformation campaigns can remove objects or people from images or create images that stage false events.

Limits to use

- We did not receive participatory disclosures from Stability AI for Stable Diffusion. This assessment is based on publicly available information, our own testing and our review process.
- Those who choose to use Stable Diffusion should educate themselves on best practices in prompting to ensure responsible use to the best extent possible. [Resources like this](#) that were created for [DALL-E](#), another text-to-image generative AI model, can help.