Stable Diffusion, Open-Access Image Generation and Disinformation

The aim of Disinfo Radar's Rapid Response Briefs is to identify new and noteworthy disinformation technologies, tactics, or narratives. Such cases may include the identification of a new technology that may have harmful uses.

Background

Consider a scenario in which a hostile actor creates a false headline, builds a story around it, and uses artificial intelligence (AI) to create an image that perfectly supports their lie. Is this a dystopian scenario, detached from reality? With the advances in text-to-image generation, this may soon become possible, as research in this domain is moving ahead rapidly, incrementally allowing for the manufacturing of realistic fake evidence.

Text-to-image generation is one of this summer's tech trends, with the potential to impact our general online experiences – for better or worse. For weeks now, thousands of AI-created motifs that look shockingly accurate have been circulating online, based on simple text prompts. These have emanated from DALL-E 2, its less sophisticated replica Craiyon, commonly known as "DALL-E mini", or the most recent market entry, Stable Diffusion, an open-access model that not only approaches DALL-E 2-in quality, but also lacks output restrictions and safeguards.

What is Text-to-Image Generation?

Text-to-image generation is reshaping the integrity of media content. Whereas previous techniques, such as deepfakes, manipulated pre-existing videos, images or audio, text-to-image creation can generate fully synthetic content, seemingly out of nowhere. Text-to-image generation tools are based on machine-learning models trained to produce realistic images from scratch, using text prompts. There are several different trained models to generate images from text descriptions, and the model landscape is constantly expanding.[1]

1. For more information on text-to-image generation models and their disinformation potential, see DRI's "What a Pixel Can Tell: Text-to-Image Generation and its Disinformation Potential" report.

Why Is Open-Access a Threat to the Misuse of This Technology?

While this technology has not yet found its way into global disinformation campaigns, partly due to its sheer novelty and partly because it has not yet reached its full technical potential, text-to-image generation threatens to become a significant component of disinformation efforts. It is likely we are entering a phase where there will be a flood of images that purposefully present false information. Matters could be made even worse if fully automatised disinformation pipelines are established. In such a scenario, disinformation actors would be able to link automated (false) text prompts to text-to-image models, to create convincing disinformation narratives. Such an automatisation of the process would replace the tedious manipulation of authentic content, on which most disinformation campaigns currently rely, and increase the complexity of disinformation strategies in the online environment.

Threats will undoubtedly grow as the restrictions to access to such technology diminishes. While there have been efforts to mitigate the toxicity or spread of disinformation by restricting access to such models and including output filters (i.e., DALL-E 2), open-source replica models are emerging that, importantly, do not include any moderation or self-regulation. As such, Stable Diffusion points towards a dangerous trend. It is one of the most powerful AI text-to-image generators developed to date, is freely available to everyone, and comes without too many output restrictions, raising many ethical questions. For example, it does not prohibit the generation of images with racist or sexually explicit content and allows for the defamation of celebrities and other public figures, who are most vulnerable to being targets of smear campaigns.

The Stable Diffusion license explicitly forbids the use of the model for such purposes, but *Stability.ai*, its service provider, cannot control the content production on a case-by-case basis.

Blurred and disproportionate images of Joe Biden, President of the United States, being arrested, generated by Stable Diffusion. The more advanced the model becomes, the better the quality of images based on incriminating text prompts will be.

Expected Threats:

- **Weaponisation in the form of non-consensual sexual imagery and fraud;**
- **The creation of misleading images of politicians and other public figures; and**
- **The reinforcement of sexualised and racial stereotypes**

How to Respond?

- AI service providers should implement a standardised "AI responsibility" code of conduct for content creation that goes beyond self-regulation. A visible watermark as the sole proof of synthetically produced imagery is insufficient for verifying and recognising synthetic content. Model developers should, therefore, be required to maintain an industry-wide, open standard for the authentication of content.

- AI service providers should work with filters to reduce the risk of biased models. While access-restricted models forbid the generation of accurate representations of people and refuse to depict public personae, the large-scale input data that feeds the algorithms is unfiltered, and often reinforces already existing stereotypes. The use of filtered subsets or purely synthetic data to train the models would be one approach to preventing the reproduction of stereotypes.

- AI service providers should introduce regular evaluation of the potential harms of text-to-image generation models prior to the release of such models. This would help ensure the more responsible and sustainable use of these models.

- AI service providers should develop a trusting and cooperative relationship with regulators to create minimum standards (i.e. AI sandbox model trials in a controlled environment).

Date: September 2022

This Rapid Response Brief was written by Lena-Maria Böswald, Digital Democracy Programme Officer, and is part of Democracy Reporting International's Disinfo Radar project funded by the German Federal Foreign Office. Its contents do not necessarily represent the position of the German Federal Foreign Office.

About Democracy Reporting International

DRI is an independent organisation dedicated to promoting democracy worldwide. We believe that people are active participants in public life, not subjects of their governments. Our work centres on analysis, reporting and capacity-building. For this, we are guided by the democratic and human rights obligations enshrined in international law. Headquartered in Berlin, DRI has offices in Lebanon, Libya, Myanmar, Pakistan, Sri Lanka, Tunisia, and Ukraine.