Worth More than 1,000 words: The Disinformation Potential of Text-to-Video Generators

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

Background

Innovations in AI-powered image creation have reached a fever pitch this year. New technologies that allow users to create realistic images from simple text prompts, such as OpenAI's DALL-E 2, Google Brain's Imagen, and Stability AI's Stable Diffusion, brought with them distinct new potential for disinformation. And now, Meta's unveiling of the world's first publicly accessible text-to-video generator – Make-A-Video, followed by the release of Google's Imagen Video, have taken this another step forward. While DRI's research has shown that text-to-image technology offers great potential as a weapon for disinformation, fully synthetic video generators pose an even greater risk. It is paramount, therefore, to understand the potential harm that may stem from this technology.

What is Text-to-Video Generation?

In a sense, text-to-video generation represents a natural progression in the field of fully synthetic media, as it is powered by the same technology. As is the case with text-to-image generators, text-to-video generation uses machine-learning models that are trained with labelled images. From there, the model is trained with unlabelled video sequences to learn how to sequence individual frames into a rapid time sequence. In other words, the resulting model can splice together multiple images at a rapid speed, creating a video in a similar way to a cartoon flipbook.

Synthetic Videos Lead Us Further to a Synthetic Reality

Both Meta and Google announced their respective models recently. Currently, they are only capable of producing five-second video clips, but they can integrate a multitude

of text prompts, allowing for more complex requests with several sentences. Based on the examples provided by Meta and Google, the models are still unable to produce hyper-realistic video sequences. But this year has shown how rapidly developments in AI-powered synthetic content are moving. Text-to-image generators have improved the quality of their outputs with each new model release. We should expect the same from synthetic videos.

Synthetic videos, like synthetic images, have a strong potential to be weaponised for spreading disinformation and hate speech, and Text-to-video generators introduce an additional risk over text-to-image generators: videos are more popular on the internet than images. Many social media platform algorithms favour and promote video content, and users prefer them. The internet is dominated by short videos and memes, with them becoming an alternative to discussing topics online, or to the political humour of traditional media.

Although, as mentioned, these generators currently only produce short video clips, the technology is bound to improve, and could potentially produce content indistinguishable from authentic video. Experts worry about the use of video memes as a mainstay of influence operations, due to their huge disinformation potential. Memes do not need to reflect reality; imagination and creativity are at the core of meme culture, meaning that the line between satire and fact can be difficult to tease out.

Text-to-video generators have the potential to inundate a video- and meme-driven internet with fully synthetic videos. As was the case with their text-to-image predecessors, many groups will be at risk of increasingly sophisticated disinformation content from the new technology. Well-known individuals, such as politicians or celebrities, could be subjected to synthetic media depicting them without their knowledge or consent. DRI's research shows that the threat will be even greater for mid-to-low level politicians or celebrities, who do not have the resources to fund public relations teams to protect them and their online presence. Likewise, charged political developments, such as wars or elections, will provide fertile ground for disinformation actors to misrepresent the facts using synthetic video. Marginalised groups may also be targeted, further impacting their ability to participate in online spaces. Finally, women's online participation might be particularly burdened by an increase of non-consensual sexual content, or revenge porn, as has been the case with deepfakes.

As observed previously with text-to-image models, the release of a text-to-video model by a major company is more than likely to be followed swiftly by competing models, with improved algorithms and less compliance with responsible AI practices. One such model, Phenaki, generates longer videos and uses more detailed text prompts, albeit at a lower resolution than the models offered by Meta and Google. As occurred with text-to-image technology, we should expect open-source models that lack moderation or self-regulation mechanisms to follow. The lack of provenance verification is only one of the risks associated with these follow-up models.

Text-to-video models may also be trained using socially biased data, which will exacerbate social stereotypes and prejudices. Producing stereotypical representations with videos generated from text prompts and reinforcing already existing overly sexualised and racial stereotypes against women and people of colour is a concern, as the models learn concepts from enormous pools of online text, images and other data that already show bias.

So far, as a strategy to mitigate the potential misuses of their models, both Meta and Google include watermarks in the video outputs, but even these can easily be clipped out. In line with responsible AI practices, Meta and Google attempt to mitigate bias and sexual violence by filtering out imagery of a pornographic nature and a lexicon of toxic words from its datasets. But there is a significant risk that follow-up models will not contain any provenance rules and will have looser safeguarding measures, similar to developments in text-to-image generation, where initial safeguards (like those imposed by OpenAI) were dropped by open-access follow-up models, such as Stable Diffusion.

What's the Threat?

- Weaponised revenge porn or non-consensual sexual video content that is longer, more realistic, and more compelling than existing fake sexual images;
- The dissemination of disinformation materials through the creation of memes and humorous video content;
- The creation of misleading videos of politicians and other public figures;
- The reinforcement of sexualised and racial stereotypes; and
- The release of new models with improved algorithms by other developers, replicating reality in video content.

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 videos 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, as well as to comply with provenance technology guidelines.
- The European Union should update its Ethics Guidelines For Trustworthy AI, known as the "Code of Practice", to include standards for all text-to-video generation models.

- AI service providers should reduce the risk of biased models, using either filtered or purely synthetic data. 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, input prompts or purely synthetic data to train the models would be one possible approach to preventing the reproduction of stereotypes.

- The EU AI Act should require AI service providers to carry out regular evaluations of the potential harms of text-to-video generation models prior to their release. This would help ensure the more responsible and sustainable use of these models.

Date: December 2022

This Rapid Response Brief was written by Jan Nicola Beyer, Digital Democracy Research Coordinator with contributions from Beatriz Almeida Saab, Digital Democracy Research Associate, and is part of DRI'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 Inte rnational

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.