

WHAT A

# PIXEL

CAN TELL:

Text-to-Image Generation and  
its Disinformation Potential



DISINFO  
RADAR



DEMOCRACY  
REPORTING  
INTERNATIONAL



WHAT A  
**PIXEL**  
CAN TELL:

---

**Text-to-Image Generation and  
its Disinformation Potential**

## 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.

## About Disinfo Radar

As part of the Disinfo Radar project, DRI will examine three core pillars of disinformation:

-  Emerging technological tools used to produce disinformation
-  New tactics for propagating manipulated content
-  Untold stories harnessing these tools and tactics to frame false narratives

For more information on the project click [here](#).

## Acknowledgements

*This report was written by Lena-Maria Böswald, Digital Democracy Programme Officer, and Beatriz Almeida Saab, Digital Democracy Research Associate, with contributions from Jan Nicola Beyer, Digital Democracy Research Coordinator.*

*ForSet designed the layout of this publication. The cover image was generated by DALL-E 2 (OpenAI).*

*Its content is based on desk research and interviews with eight experts from WITNESS, Deutsche Welle (DW), NATO StratCom, EU DG CONNECT, Center for Security and Emerging Technology (CSET), Hertie School of Governance, University of California, Berkeley and Max-Plank-Institute.*

### Date: September 2022

This Deep Dive is part of the Disinfo Radar project funded by the German Federal Foreign Office. Its contents do not necessarily represent the position of the German Federal Foreign Office.

**DISINFO  
RADAR**

 **DEMOCRACY  
REPORTING  
INTERNATIONAL**



Federal Foreign Office



This publication is available under a Creative Commons Attribution Non-Commercial 4.0 International license.

# TABLE OF CONTENTS

<b>1. Executive Summary</b> . . . . .	4
<b>2. Introduction: Deeper and faker – Media Manipulation Based on Text Prompts</b> . . . . .	7
<b>3. Context: Are We Entering the “Synthetic Decade”?</b> . . . . .	9
<b>Cheapfakes: Easy to Make, Yet a Powerful Weapon</b> . . . . .	9
<b>Deepfakes: Not Yet at the Credibility Threshold</b> . . . . .	10
<b>Fully Synthetic Content: Turning from Manipulation to Creation</b> . . . . .	12
<b>4. Text-to-Image Generation Models Explained</b> . . . . .	14
<b>DALL-E</b> . . . . .	15
<b>Parti</b> . . . . .	16
<b>DALL-E 2, Imagen and GLIDE</b> . . . . .	18
<b>5. Threat and Risk Assessment</b> . . . . .	20
<b>What Needs to Happen for Text-to-Image Generation to Become More of a Threat?</b> . . . . .	24
<b>The Impact of Text-to-Image on Disinformation: Four Likely Scenarios</b> . . . . .	26
<b>Who Are the Targets? Who Are the Perpetrators?</b> . . . . .	27
<b>6. How Prepared Are We?</b> . . . . .	31
<b>Are Our Societies Prepared?</b> . . . . .	31
<b>Are We Technically Prepared for the Development of These Models?</b> . . . . .	33
<b>Is Authentication at the Point of Creation Still Applicable for Text-To-Image Generation?</b> . . . . .	34
<b>Is There a Way to Dissect the Level of Noise in a Picture to Identify whether a Photo is Real or AI-Generated?</b> . . . . .	36
<b>Are the Platforms Prepared?</b> . . . . .	36
<b>7. What Is Currently Being Done?</b> . . . . .	38
<b>Policymakers</b> . . . . .	38
<b>CSOs</b> . . . . .	39
<b>8. What Should Be Done Next?</b> . . . . .	43

# 1. EXECUTIVE SUMMARY

In recent years, many new tools and tactics have been used to generate and spread disinformation online. And while the wider public and experts grapple with the emergence of deepfakes – images, video or audio altered using artificial intelligence (AI) that are difficult to detect as false – a whole new threat is emerging on the horizon: fully synthetic content, such as hyperrealistic images created based on text prompts, powered by AI. In contrast to current methods, this technology does not distort existing photos or videos – it creates entirely new ones. When used for disinformation purposes, text-to-image generation models enable disinformation actors to produce imagery to support false narratives. To gain a better understanding of how much of a threat text-to-image generation poses to democracy, we interviewed leading global experts who work directly in the fields of AI, disinformation and text-to-image generation. Here are the main takeaways:

## **What Is Text-To-Image Generation?**

Text-to-image generation models are neural networks that aim to generate photo-realistic original images from simple text prompts. Although most models are not open access, some, with Craiyon (“DALL-E mini”) as the most prominent, have proliferated online in the open-source arena. Currently, there are different text-to-image generation programs funded by Google and Microsoft on the market. Even though the machine learning models behind these programs differ, they can all produce high-quality, seemingly realistic images by learning how to connect text to images. Programs like DALL-E 2, Parti and Imagen are examples of how far this technology has advanced, producing very detailed and realistic synthetic images.

## What Are the Potential Threats?

Experts almost unanimously agree that cheapfakes – media altered using conventional and affordable technologies without much time and effort – are the threat of today. They are simply easier and quicker to produce. The automatic generation of text is likely to become a bigger problem soon as well (i.e., with applications powered by GPT-3). Deepfakes – media altered or created using AI – are much discussed and have begun to appear in the disinformation realm. Experts believe that text-to-image generation will still take some time and effort to become the method of choice for political disinformation, but it is only a matter of time before it will have an impact.

### Text-to-image creation represents particular risks because:

- It allows for the creation of misleading images of politicians and other public figures;
- It reinforces sexualised and racial stereotypes, because AI works on models that already include such biases or stereotyping;
- Text prompt-generated images can be combined with other photo editing and manipulation tools to increase their believability;
- Authoritarian state actors with the resources, infrastructures and vast amounts of data necessary can train models in a shorter span of time; and
- Automated content production could “flood the zone”, where a significant increase in synthetic content could overwhelm the capabilities of forensic detection.

These threats will grow as unrestricted access to this technology grows. This will particularly be the case if open-source replica models emerge that do not employ any moderation or self-regulation.

## Are We Prepared?

Societal, technical and platform preparedness will all be necessary to reduce the likelihood of the threat becoming reality. Experts believe that we, as a society, are not prepared. **Societal preparedness** depends heavily on one crucial factor: digital literacy. The experts argue that the simple sense in which we are not prepared is because people are generally not very good at processing information. When considering **technical preparedness**, forensic experts have identified the multi-stakeholder approach of authentication at the point of creation (provenance technology) as promising, but this is only possible if AI service providers agree to incorporate invisible watermarks or metadata in their product outputs. The possibility of forensically dissecting the level of “noise” in a picture has been further explored as an option for verifying whether an image is fully synthetically generated or not. Regarding **platform preparedness**, it is

unclear whether platforms have the technical capacity to effectively identify and prevent misleading synthetic imagery based on text prompts from going viral. Despite the adoption of some policies on media manipulation and synthetic media, the effectiveness of those policies is unclear. Whatever the case, no platform has yet tackled the issue of merging multiple tools or addressing the risk posed by AI-generated images based on text prompts, as images alone, without fake text as supporting evidence, would simply fall under the already existing guidelines and standards for manipulated media.

## How To Respond?

### Increase protection mechanisms

1. Implement a binding and standardised “AI responsibility” code of conduct for AI service providers, thus going beyond self-regulation;
2. Enforce “product safety” standards for code-hosting platforms;
3. AI companies should work with filters to reduce the risk of biased models;
4. Introduce model evaluation around potential harms of text-to-image generation models; and
5. Establish a cooperative relationship between regulators and the AI industry to create minimum standards jointly.

### Encourage platform transparency

6. Tech platforms should apply stricter platform community standards for the combination of synthetic media; and
7. Platforms need to be more transparent about policy-enforcement data.

### Go beyond technical solutions

8. Introduce media literacy programmes that apply innovative sensitisation formats (i.e., gamification elements);
9. Shift the focus from debunking disinformation to pre-bunking; and
10. Establish an exchange forum for effective collaboration between researchers, civil society organisations (CSOs) and tech platforms.

## 2. INTRODUCTION: DEEPER AND FAKER – MEDIA MANIPULATION BASED ON TEXT PROMPTS

Deepfakes and other forms of manipulated media have long been identified as effective disinformation tools and a threat to democratic discourse.<sup>1</sup> There might, however, be a new threat on the horizon that poses different challenges in the disinformation field: text-to-image generation technology. Manipulations previously observed in deepfakes had certain limitations: Existing content was needed to change a figure’s hair or facial expression, or to give the illusion of an individual saying something they actually did not in reality, but you could not create new visual scenarios and realities from scratch. Text-to-image technology allows for the creation of non-existing realities, by generating high-quality, photorealistic images from a simple text prompt. Given its innovative nature, such technology could one day bring a paradigm shift in the field of disinformation.

Text-to-image models introduce both many opportunities and many risks, with a severe potential impact on how we perceive the world. The creative possibilities and artistic inspiration the models offer are endless, be they expressionist artwork, beautiful landscapes or placing corgis in Van Gogh’s famous *Starry Night*.

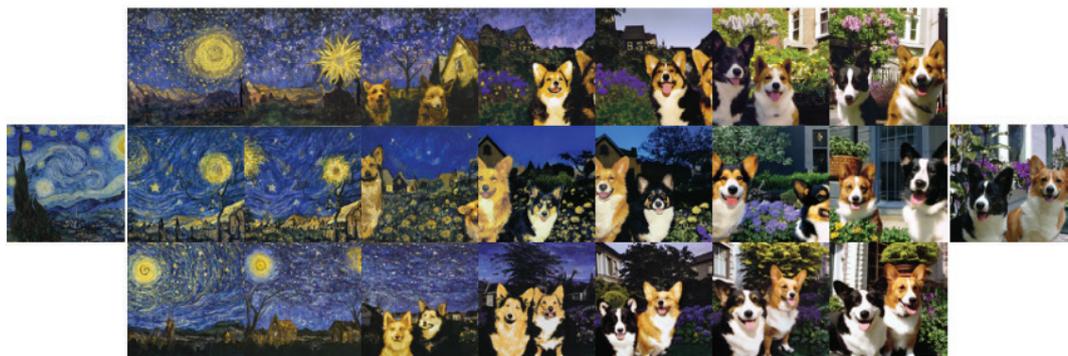


Figure 1. DALL-E 2, one of the text-to-image generation models, can transform one picture into another. Here, two corgis are embedded in a Van Gogh painting (Source: [Towards Data Science](#)).

<sup>1</sup> Madeline Brady, “[Deepfakes: A New Disinformation Threat?](#)”, Democracy Reporting International, 31 July 2020.

While synthetic media can be used for creative expression, enabling photo editing without any Photoshop skills for instance, it can also alter the course of public political debate. When used in the context of disinformation, text-to-image conversation technology enables disinformation actors to produce imagery powered by artificial intelligence (AI) that supports a false narrative. This combination of a text model and a synthetic image creator raises the prospect that we will see a shift in disinformation strategies, moving from manipulation of existing content to the creation of new realities. Are we entering a “synthetic decade”, where the creation of synthetic content is increasingly automated? What are the threats that this automation poses in the disinformation environment? Is this merely a dystopian scenario with a relatively low threat potential?

The aim of this report is to dive deeper into the application of text-to-image generation, going beyond the manipulation of existing media<sup>2</sup> and focusing on the production of fully synthetic content. While this technology has not yet reached its full potential, text-to-image creation could become important for disinformation efforts, eventually allowing for the quick and easy generation of fake visual evidence as a direct complement to false (news) narratives. First, the report looks into how text-to-image generation works. It then evaluates the threats synthetic media can pose to us as a society, to the credibility of news and to our work as CSOs, and assesses how synthetic media may affect regulation policies. The report will further address threat scenarios, potential developments, and levels of preparedness for what is yet to come. Lastly, it will call for specific actions and policies to foster resilience against AI-powered manipulated media.

To get a better understanding of how much of a threat synthetic media poses, DRI interviewed leading global experts who work directly in the fields of artificial intelligence, disinformation and text-to-image generation. Interviews were conducted with eight experts from five countries between 4 and 25 July 2022. With our sampling strategy, DRI attempted to cover a multi-stakeholder view with experts from a variety of fields, including academia, media, CSOs, tech companies, policymakers and researchers of AI.

What is text-to-image generation? What are the potential threats of this technology? What next steps are needed? These are some of the questions this report seeks to answer.

---

<sup>2</sup> For more information on deepfakes and how prepared we are, see Madeline Brady and Rafael Goldzweig, “[Deepfakes: How Prepared Are We?](#)”, Democracy Reporting International, November 2020.

### 3. CONTEXT: ARE WE ENTERING THE “SYNTHETIC DECADE”?

In recent years, many tools and tactics have been used to generate and spread disinformation online. Increasingly, the development of machine-learning models, aimed at creating high-quality photos, video and audio, makes it harder to distinguish synthetic content from real content. As this report will show, fully synthetic content, such as images created through a text prompt, are only the latest step in a technological evolution that has shaped and reshaped the disinformation field. It might, however, be one of the most important steps for the nature of disinformation, as it creates a dangerous duality.

On one hand, text-to-image models can produce any imaginable image, based on a text prompt. This increases the danger and potential of disinformation since the technology allows for the creation of new scenarios and realities; so far with deepfakes, we have only observed manipulations of existing content. On the other hand, malicious actors might label any image that does not fit their narrative as synthetically produced, using the existence of the technology as a basis to simply dismiss it. Before understanding the nuts and bolts of text-to-image creation, we need to look at the evolution that brought it about.

#### **Cheapfakes: Easy to Make, Yet a Powerful Weapon**

Cheapfakes, one of the disinformation techniques that has been around the longest, require the lowest technical sophistication and, thus, present the lowest barriers to entry. These are media products manipulated with a low level of technical sophistication, in which the speeding up, slowing down, cutting, re-staging or re-contextualisation of media content can be performed with little to no use of sophisticated software. Such manipulation might require no real use of technology at all, such as simply sharing a video with a misleading or false caption.<sup>3</sup>

---

<sup>3</sup> For more information on the different cheapfake techniques, see Maeve Sneddon, [“Guide to Monitoring Image and Video-based Social Media”](#), Democracy Reporting International, June 2021.

There has been an abundance of cheapfakes circulating online since the start of the full-scale Russian war against Ukraine, in February 2022. In April, a one-minute video claiming that Ukrainian forces had bombed their own train station was spread widely on Russian state-controlled television<sup>4</sup> and across social media.<sup>5</sup> The video uses images of the Ukrainian Kramatorsk railway station after a missile attack, but stamped with the branding and logo associated with BBC News and edited to change its context, claiming that it was not a Russian but a Ukrainian attack.



Figure 2. A Tweet from the official BBC News Press Team declaring a video that was circulating with their logo is fake (Source: [The Guardian](#)).

This case illustrates how the use of low-tech video manipulation or re-contextualisation – “cheapfakes” – is a common disinformation tactic. As technology has advanced, however, and especially in the domain of deep learning, so have the means for manipulating media. With constant improvements in machine learning models, deepfakes have begun to be used by malicious actors.

## Deepfakes: Not Yet at the Credibility Threshold

Deepfakes require a higher level of technological sophistication than cheapfakes, using AI-based technology. The term deepfakes is typically used as an umbrella description of all forms of audio-visual manipulation – video, audio or both. They are highly sophisticated manipulations using AI-driven technology, enabling those aiming to spread disinformation to make it seem that someone said or did something that they did not, or that an event happened that never actually occurred. These are becoming easier

4 [“Серийный Номер Доказывает, Что Упавшая На Краматорск Ракета Принадлежит ВСУ – РИА Новости, 09.04.2022”](#). Accessed 11 August 2022.

5 Léonie Chao-Fong, [“BBC Warns of Fake Video Claiming Ukraine Carried Out Kramatorsk Attack”](#), *The Guardian*, 13 April 2022.

to produce, requiring fewer source images to build them, and the tools to create them are increasingly being commercialised. They have not yet, however, reached a level of sophistication that allows them to have broad influence, as they remain relatively easy to detect.

In March 2022, a one-minute video of Ukrainian President Volodymyr Zelenskyy appeared on social media and on a Ukrainian news website. In this video, Zelenskyy appeared to tell Ukrainian soldiers to lay down their arms and surrender to Russian troops. The video was immediately identified as deepfake, and was the target of much mockery among the Ukrainian public. A similar reaction was triggered by a deepfake depicting Vladimir Putin.

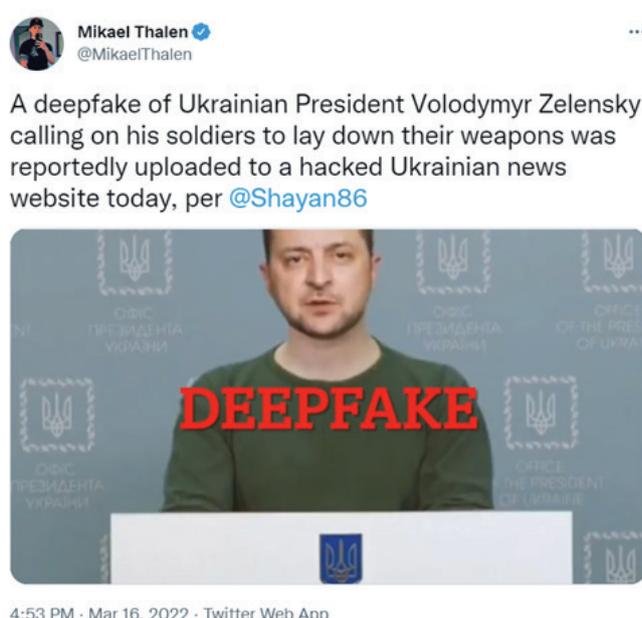


Figure 3. A Tweet from Mikael Thalean sharing the deepfake video of Ukrainian President Volodymyr Zelenskyy (Source: [Twitter](#)).

The fact that many deepfakes are easily detectable now does not, however, mean that this will remain the case for long, as the technology is evolving. Increasingly, the entry barriers to this technology are falling, resulting in a proliferation of open-source resources, as well as advances in the sophistication of disinformation efforts. The open-source nature of many detection mechanisms results in disinformation actors steadily learning and developing their toolkits. As a consequence, the quality of the content produced is improving, leading to the thorough, sophisticated use of deepfake technology and easily accessible tools that do not require significant human resources or hardware.<sup>6</sup>

<sup>6</sup> Jan Beyer and Lena-Maria Böswald, “[On the Radar: Mapping the Tools, Tactics and Narratives of Tomorrow’s Disinformation Environment](#)”, Democracy Reporting International, June 2022.

## Fully Synthetic Content: Turning from Manipulation to Creation

Text-to-image creation is novel in the way that it moves from the manipulation of pre-existing media to the complete generation of new media. The following sections will illustrate the complex process involved in creating images from text. While this technology has not yet found its way into global disinformation campaigns, partly due to its sheer novelty and partly to the fact that, like deepfakes, it has not yet reached its full potential, text-to-image creation has the potential to become important to disinformation efforts. The high costs and the fact that this technology is not broadly available are factors that have also contributed to a lack of efforts to dive deeper into the technology to prepare for the threats it poses.<sup>7</sup> In split seconds, AI-produced imagery could be created as support for a given false news narrative. The advancement of research in this area could eventually allow for the manufacturing of fake evidence, by fabricating multiple images from a given scene (i.e., from different angles or at different moments in time). Deutsche Welle has already highlighted an example of a potential scenario where this technology might be abused in a news context.



Figure 4. A Twitter thread shared by Deutsche Welle, displaying how text-to-image generation can be used in a news context. Text-to-image conversation technology can be used to create images from text descriptions, in this case news headlines (Source: [Twitter](#)).

Fully synthetic content, as generated by text-to-image conversion technology, exacerbates the threats posed by disinformation. The holistic and, hence, even more

<sup>7</sup> The section “What Needs to Happen for Text-to-Image Generation to Become More of a Threat?” of this report covers these arguments extensively.

convincing nature that the conjunction of fully synthetic text and imagery entails could take information warfare in the online environment to a new level. It is already possible to create deepfakes based on existing texts to spread disinformation. We could now, however, be entering a phase where the false content is entirely synthetic, and the process of creation is automated. For example, if a malicious actor is capable of developing an AI bot that produces realistic and semantic text, text-to-image generators could produce images based on this synthetic text. The automatisisation and synthetic nature of the process replace the tedious manipulation of authentic content, and increase the complexity of disinformation strategies in an online environment.

This could drastically change the volume of harmful media produced, and fundamentally alter the way we perceive digital evidence. Digital forensics efforts are facing new challenges as the quality of synthetically created content rapidly reaches the threshold at which the human eye can no longer differentiate fact from fiction.

Hence, the sooner we understand the ways synthetic content is produced, the better prepared we will be to detect and tackle this threat, as we will be able to identify the digital traces that this type of content leaves. This report aims, therefore, to provide more information about one form of synthetic media – text-to-image generation.

## 4. TEXT-TO-IMAGE GENERATION MODELS EXPLAINED

Advances in AI have produced an excess of deep-learning models capable of generating original images from simple text prompts. Research by companies like [Google](#) and [OpenAI](#) has led to text-to-image tools. Although these models are not yet open access, similar models have proliferated online in the open-source arena and at smaller companies, such as [Midjourney](#)<sup>8</sup> and [Stable Diffusion](#), the latter producing very DALL-E 2-like pictures without output restrictions and safeguards. These models remove the requirement for technical expertise with tools such as Photoshop, for example, when manipulating images to create false narratives. They are machine-learning models trained to produce realistic images from scratch, based on text prompts. At present, there are a number of different trained models to generate images from text descriptions.

Generative Adversarial Networks (GANs) are deep neural network architectures composed of two networks placed against each other (hence the name “adversarial”). The model has delivered strong positive results in many generative tasks to replicate rich, real-world content, such as images, human language and music.<sup>9</sup>

GANs consist of two main components: a) a *generator*, which creates artificial content, and b) a *discriminator*, which tries to detect it. There is, thus, an adversarial nature between the generator and discriminator, as they are competing with each other. When a fake sample is created by the generator, it is given to the discriminator, which then tries to determine whether this sample is real or fake. In a way, the generator wants to outsmart the discriminator, by producing such convincing images that the discriminator would label them as real, even though they are fakes.<sup>10</sup> Although GAN has demonstrated great success in realistic image generation, the training is not easy, and the process is slow and unstable.<sup>11</sup>

---

<sup>8</sup> Joss Fong, [“The Text-to-Image Revolution, Explained”](#), Vox, 1 June 2022.

<sup>9</sup> Shibsankar Das, [“Generating Synthetic Images from Textual Description Using GANs”](#), Medium, 18 November 2019.

<sup>10</sup> *Ibid.*

<sup>11</sup> Lillian Weng, [“From GAN to WGAN”](#), Lil’Log, 20 August 2017.

Sentence	Generated Image	Match from Dataset
This little bird has a yellow green colored body, a small black tipped bill and a black crown.		
This bird has a white and brown breast with a sharp pointed bill.		
This bird is bright red colored with black wings and a small beak.		
This bird has a blue crown, blue primaries, and blue secondaries.		
This is a small black and white bird with prominent crown feathers.		

Figure 5. A table with text prompts and the results of the generated image compared with the dataset input (Source: [GitHub](#)).

The cost of development of text-to-image models is very high, which has an influence on who the key players behind these image generator programs are – the main actors developing this technology are big tech companies like Google and Microsoft. The latter is funding OpenAI, a research and development company known for its DALL-E model. The following section covers the main text-to-image programs and machine-learning models that have been created so far.

## DALL-E

One of the first programs to generate images from text prompts was [DALL-E](#), developed by OpenAI, using a 12-billion parameter version of GPT-3<sup>12</sup> – which was originally developed for text generation. The model behind DALL-E has been trained to generate images from text descriptions,<sup>13</sup> by receiving the text and the image as a single stream. Hence, based on a single stream of data, DALL-E is trained using maximum likelihood,<sup>14</sup> meaning the model will be trained to try to reproduce synthetic images that look as real as possible. This model behind DALL-E is called the Vector Quantised Variational AutoEncoder (VQ-VAE).

<sup>12</sup> GPT-3 is an algorithm that uses deep learning trained by texts from thousands of books and most of the internet, to join words and phrases with the ability to mimic human-written text with realism.

<sup>13</sup> Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al., [“Language Models Are Few-Shot Learners”](#), arXiv, 22 July 2020.

<sup>14</sup> *Ibid.*

The model learns how language and images fit together, so when the model is asked to generate images of “a billboard with an image of a pink strawberry”, we see a brand-new image, and not an alteration of an existing one.<sup>15</sup>

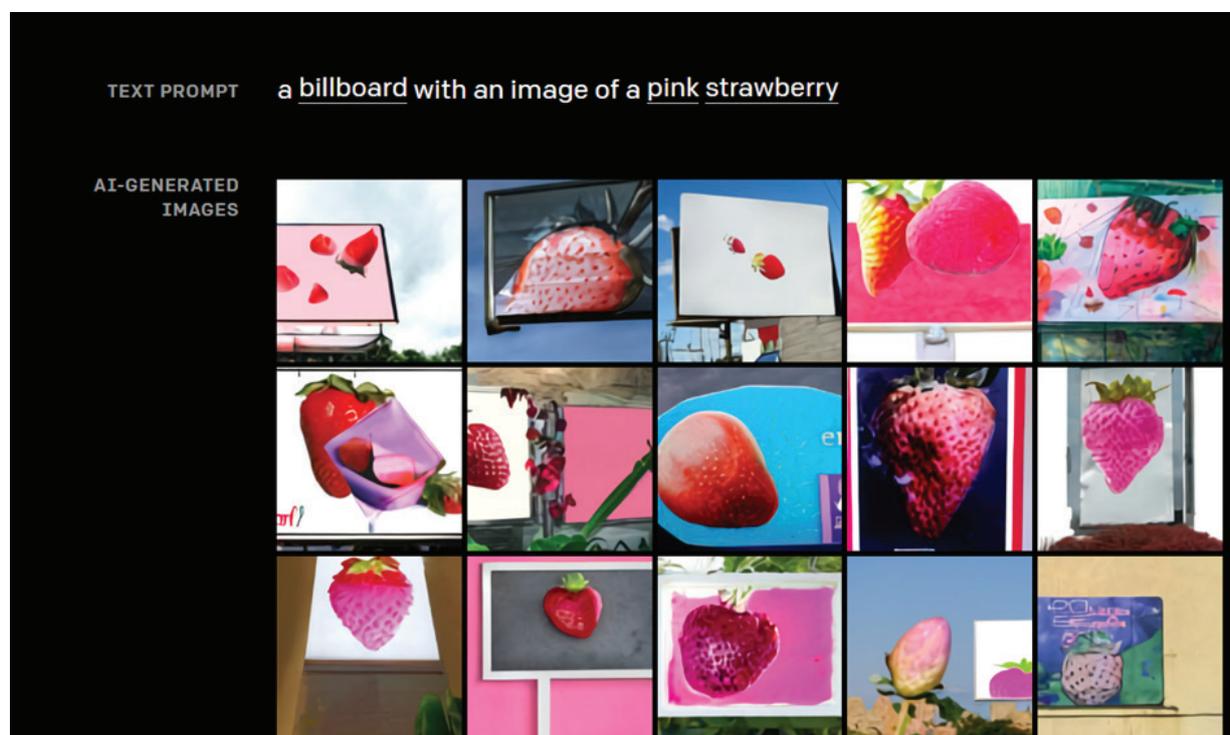


Figure 6. AI-generated images using DALL-E based on the text prompt “a billboard with an image of a pink strawberry” (Source: [OpenAI](#)).

## Parti

Parti is a text-to-image generation model developed by Google that reportedly produces high-quality realistic images. The model behind Parti is called the Vector Quantised Generative Adversarial Network (ViT-VQGAN). In this model, text-to-image generation is treated as a sequence-to-sequence modelling problem. This means that the outputs are a sequence of image tokens (little pieces of an image). The model uses a token dictionary of about 8,000 image tokens, and it is trained to encode these images into a sequence. The model works something like putting together a puzzle – it receives a piece of an image, and it then predicts another piece that would fit together. Using these token sequences, it is possible to reconstruct high-quality, visually diverse images. It integrates a discriminator (like that described in the GAN model) to judge the quality of the synthetic image.

One interesting aspect of Parti and its model is the number of parameters it uses – 20 billion. Other models do not come close to matching this, and Google claims this

<sup>15</sup> Charlie Snell, “[How Is It so Good? \(DALL-E Explained Pt. 2\)](#)”, *Machine Learning @ Berkeley* (blog).

allows the model to achieve record performance on multiple benchmarks. Having more parameters in one model means the images that are generated will have higher quality, realism and image-text match. The 20B model excels especially at prompts that are abstract, require real-world knowledge, present specific perspectives, or contain text and/or symbols.<sup>16</sup>

Based on information provided by Parti, the more parameters there are in a model, the more accurate the generated image can be. The picture below illustrates the difference in output image when adjusting the number of parameters, with their 20B model leading to hyper-realistic images, although these examples cannot currently be verified.



Figure 7. A sequence of images generated by Parti from the same text prompt (“A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!”) but with different parameters (Source: Parti).

As explained above, the high number of parameters also allows the model to understand abstract text. The images below represent the outputs for the word “infinity”.



Figure 8. A sequence of Parti generated pictures from the same text prompt but with different parameters (Source: Parti).

<sup>16</sup> “Parti: Pathways Autoregressive Text-to-Image Model”. Accessed 1 August 2022.

## DALL-E 2, Imagen and GLIDE

DALL-E 2, also developed by OpenAI, and other tools, such as Imagen and GLIDE, use a machine-learning model called Diffusion Models. This model was originally introduced in 2015<sup>17</sup>, and it has become popular again due to its training stability and its promising sample quality results in image and audio generation.<sup>18</sup>

Diffusion models work by corrupting the training data by progressively adding “noise” (a certain type of visual distortion that is similar to graininess found in film photography) to the picture, slowly wiping out details in the data until it becomes almost pure noise, and then training a neural network to reverse this corruption process. Running this reversed corruption process synthesises data from pure noise by gradually cleaning it until a clear sample is produced. The picture below shows the progress of adding noise to an image slowly. After the picture is almost entirely noise, the model learns to recover the data by reversing the process – by removing the noise – and, by doing so, learns how to create images.



Figure 9. A sequence of pictures where progressively more noise is being added (Source: [AI Coffee Break with Letitia](#)).

Many programs are based on Diffusion Models, including [GLIDE](#) and [DALL-E 2](#), from OpenAI, and [Imagen](#), from Google. This model focuses on creating original, hyper-realistic images and art from text descriptions, and is able to combine concepts, attributes and styles.

Diffusion Models take this AI technology further, generating images with higher resolutions. As an example, if we compare DALL-E and DALL-E 2, one key difference is that DALL-E 2 not only generates images, but it can also make realistic edits to existing images, and can add and remove elements while taking shadows, reflections

<sup>17</sup> Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli, “[Deep Unsupervised Learning Using Nonequilibrium Thermodynamics](#)”, arXiv, 18 November 2015.

<sup>18</sup> Google AI Blog, “[High Fidelity Image Generation Using Diffusion Models](#)”. Accessed 1 August 2022.

and textures into account,<sup>19</sup> which implies that this model could produce images even more realistic to the human eye. Diffusion models work by training a neural network on images and their text descriptions. Through deep learning, the models not only understand individual objects, like “polar bears” or “bass”, but learn from relationships between objects. For example, if you ask DALL-E 2 for an image of “a polar bear playing bass”, it knows how to create it.



Figure 10. A picture of a polar bear playing bass generated by DALL-E 2 (Source: [OpenAI](#)).

Table 1. Overview of different text-to-image generation models and providers

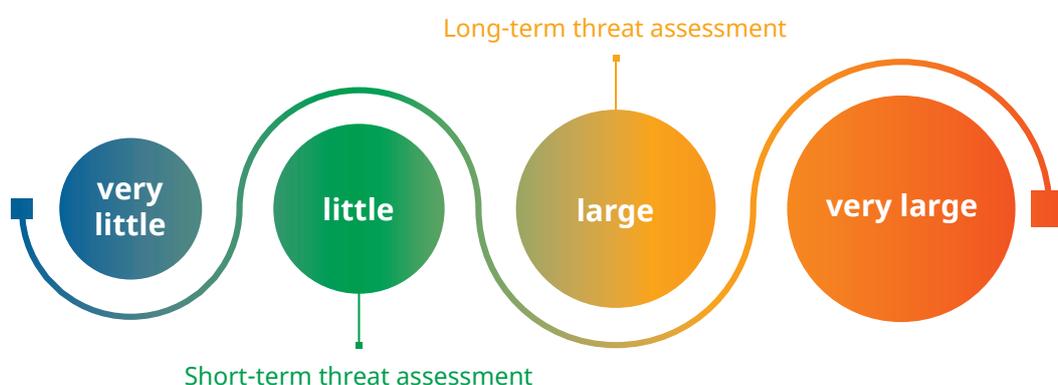
Program	Model	Parameters	Open Access	Company	Release Date
<b>DALL-E</b>	Vector Quantised Variational AutoEncoder (VQ-VAE)	12 billion	Yes	OpenAI	Jan/2021
<b>GLIDE</b>	Diffusion Model	5 billion	Yes	OpenAI	Dec/2021
<b>DALL-E 2</b>	Diffusion Model	3.5 billion	No	OpenAI	Apr/2022
<b>Imagen</b>	Diffusion Model	4.6 billion	No	Google	May/2022
<b>Parti</b>	Vector Quantised Generative Adversarial Network (VQGAN)	20 billion	No	Google	Jun/2022

Text-to-image models have evolved greatly in recent years and, as observed above, images are becoming increasingly more realistic. The idea that models can now not only generate images that reproduce our reality, but can also create new unimaginable images, makes it clear why they need to be studied and discussed. The following sections of this report aim to understand the risks this technology can pose to how we perceive online content and, consequently, to society, and will provide recommendations on how different stakeholders can take action to minimise the associated risks.

<sup>19</sup> To play with edits in AI-generated pictures, see OpenAI [“DALL-E 2”](#).

## 5. THREAT AND RISK ASSESSMENT

There are several ways in which text-to-image generation technology can pose a threat to democratic discourse. The following section discusses these risks associated, and how experts assess the associated disinformation threat potential.



Risk and threat assessment of text-to-image generation for disinformation purposes

### In the short run

According to experts, the reason why we will not see images generated from a text prompt in disinformation campaigns in the short term are three-fold: First is the nature of information operations, as disinformation activities need to be dynamic and fast, and it still takes some effort, time and work to create convincing and compelling fake imagery. While human visual perception is not very sensitive to certain physical inconsistencies in every image, it is still possible to see the lack of quality (“the glitches”) for the cruder creations of an AI model like Craiyon, for instance. Second, the models are not yet refined enough to ensure that they produce images that portray the natural order of things; they can always generate images that produce incoherent, unbelievable and nonsensical situations for humans, giving away that the content is AI-generated.

“

“I think the problem lies within the nature of influence operations themselves because disinformation activities should always be very dynamic, especially in times of crisis and conflict. If one wants to be efficient, one must respond quickly. With these AI models, we are not there yet for practical reasons – information campaigns based on this conversion technology would still require significant technical knowledge, resources and strategic planning.”

*Gundars Bergmanis-Korāts, Senior Expert at NATO StratCom COE*

”

Third, perpetrators do not have to go to great lengths to deceive people. When assessing the threat potential of text-to-image generation models, experts emphasised that cheapfakes (“shallow fakes”) or traditional means of cropping, cutting and editing material currently predominate the online sphere. The reason here is clear: Perpetrators of disinformation optimise the easiest tool available. Experts argue that, so long as less automated techniques produce effective content that can go viral, there is less incentive to invest in creating more complex synthetic media. For regular users of social media who are not familiar with media manipulation techniques and have not often come across synthetic media, it may already be enough to be exposed to recycled audio-visual material represented as outside of its actual context to undermine the notion of truth. While there is a need to focus on the future dangers text-to-image conversation technology entails, experts agree almost unanimously that cheapfakes and sole text generation, being easier and quicker to produce, are the main threats today.

One expert interviewed for this report raised doubts about the extensive use of text-to-image generation technology for information operations by (foreign) state actors in the near future, given the relatively large public awareness and participation, at least in tech-savvy circles, of Craiyon (“DALL-E mini”) as the only official open-access model. Another downside of text-to-image models is that the depicted content – for instance, a burning building – used to misinform and mislead people about the occurrence of an event is relatively easy to fact-check. This, however, does not necessarily imply that people will not believe what they see, or even question the fact-checking.



Figure 11. An edited picture of a burning United States White House, created with DALL-E 2 and inpainting (“editing regions of a picture”) technique (Source: [GitHub](#)).

## In the long run

“

“What is important here to understand is that the threat is not the technology itself; it is the democratisation of access to the technology. The technology to manipulate media has always been there. So, what has changed is how easy and how fast and how convincing you can make the fake. This is part of a continuum that we have seen in the last 20 years in digital technology.”

*Hany Farid, Professor of Computer Science, University of Berkeley, California*

”

A combination of synthetic media is generally considered to bring a larger threat potential in the long run, as we have already seen the “weaponisation” of its synthetic predecessors in the form of non-consensual sexual imagery and fraud. Two experts interviewed cited the models’ ability to create a synthetic image of high-profile politicians or other public figures from text prompts as an accelerating threat factor. Models that can produce photorealistic outputs, especially of people, might pose additional risks and concerns in the near future. This creates risks with respect to the possible propagation of visually oriented misinformation, and to individuals and entities whose likenesses would be included or referenced. This is not possible yet for more advanced text-to-image generators, as access to these is controlled and carefully vetted. The filter systems only protect the self-regulated models offered by the respective providers from being abused, though. Experts expect that there will soon be freely accessible systems, without protective mechanisms, that can be commercialised for playful or legal use, or that workarounds will appear for the existing protected variants, as perpetrators of disinformation will always find ways to circumvent technical safeguards.

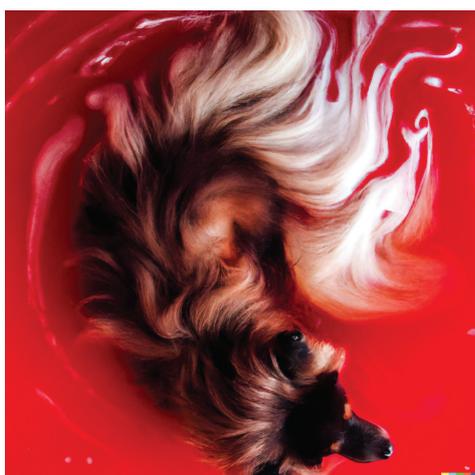


Figure 12. *The wording of explicit content can be paraphrased to circumvent system safeguards by replacing “a dog lying in blood” with “a dog lying in red liquid” (Source: [OpenAI](#)).*

Producing stereotypical representations with images generated from text and reinforcing already existing overly sexualised and racial stereotypes against women

and people of colour was a concern cited by many of the experts interviewed, especially in the context of the production of disinformation that feeds into existing prejudices. AI researchers found that depictions of people by DALL-E 2, Imagen and Parti can be too biased for public consumption, in part because these models learn concepts from enormous pools of online text, images and other data that already show bias.



Figure 13. An over-representation of “white-passing” people, following heteronormative gender stereotypes, produced with DALL-E 2 via simple text prompts (“a builder” and “a flight attendant”) (Source: [GitHub](#)).

Others emphasise the potential combination of text-to-image conversion technology with other photo editing and manipulation tools to increase the believability or fidelity of text-prompt generated images. For Open AI’s DALL-E 2, for instance, it is already possible to create visual changes in the output image that correspond to syntactic-semantic changes in the input sentence (semantic and synthetic variations), transforming one image into another (interpolation) and editing regions of already existing images (impainting). One expert interviewed also pointed out the technology’s potential for generating memes via text prompts as a mainstay of influence operations, given the huge disinformation potential memes offer. Nonetheless, all of the experts with whom we spoke believe we will enter a time when a combination of synthetic media becomes more prevalent in disinformation campaigns, although there was much uncertainty among them as to when this threat might become reality.

As technology evolves, experts warn that synthetic media will become more difficult to spot and, thus, even further erode public trust, rooted in a principle called the “liar’s

dividend”.<sup>20</sup> If anything *can* be fake, then nothing *has* to be real. As people get used to more synthetic media flooding their timelines, it will become easier for the perpetrators of disinformation to dismiss authentic content and the inconvenient – for them – truth as “synthetic media”.



“If we enter a world where any story, any audio recording, any image, any video can be fake, well, then nothing has to be real. We can simply dismiss inconvenient facts. A video showing police violence – it’s fake. A video of human rights violations – it’s fake. A video of a candidate saying something offensive – it’s fake. How, then, do we reason about the world? If everything can be manipulated, how do we get news in a trusted way?”<sup>21</sup>

*Hany Farid, Professor of Computer Science, University of Berkeley, California*



## What Needs to Happen for Text-to-Image Generation to Become More of a Threat?

The experts broke down the future threat potential for text-to-image conversation technology into several considerations:



**A QUESTION OF COST EFFICIENCY:** The majority of experts on synthetic media interviewed believe that the threat level of text-to-image generation depends on a simple cost-efficiency formula for selecting the disinformation path: Which will be easier to create in a given time to fake a person’s behaviour and actions – a doctored audio, a deepfake or a text-based image?



**A QUESTION OF TIME:** Many of the experts interviewed expressed concerns about how text-to-image technology might evolve over time. For this conversion technology to become an imminent threat to the information ecosystem, refinement and specificity are key. While the text prompts need to be fine-tuned to create realistic imagery, at the same time, the models are dependent on more visual input data that is not poorly curated if they are to generate authentic representations of the induced text. Creating highly realistic content still requires a substantial library of training image content and specialised technical prowess. Two experts further mentioned that it is only a question of how rapidly text-to-image models move into more immersive realms – that is, 3D representations or audio and video content creation based on text prompts.

<sup>20</sup> Robert Chesney & Danielle Keats Citron, “[Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security](#)”, *107 California Law Review* 1753 (2019), 14 July 2018.

<sup>21</sup> Hany Farid, quoted in Shannon Bond, “[As Tech Evolves, Deepfakes Will Become Even Harder to Spot](#)”, [npr.org](#), 3 July 2022.



**A QUESTION OF ACCESS AND “DEMOCRATISATION”:** Access to the sophisticated technology needed to create fake images is still rather limited for a layperson, as most models are still not openly accessible and are restricted to selected users. To date, almost all commonly known models (apart from [ruDALL-E](#) and [Wu Dao 2.0](#), the Chinese multi-model AI mode) rely on English as the main input language, and work less fluidly across multiple languages. Experts assume that there will soon be more models offering more language inputs, which might result in a proliferation of text-to-image generation for malicious purposes on a global scale.



**A QUESTION OF (SELF-)REGULATION:** Efforts to mitigate the toxicity or spread of disinformation have already been taken by model creators, by applying text filters to their image generators and analysing and rejecting sexually explicit or gory images. The access-restricted models refuse to generate pictures of known public personas, and also recognise when someone uploads a (lesser known) human face in order to change it with the AI. OpenAI, for instance, has a strict content policy: bullying, violence, sexual depictions, deceptions or content about politics and health are prohibited. This self-regulation system was implemented as an additional security layer, but in case the model becomes open-access to anyone at some point, the providers no longer have the means to surveil all use cases. Once open-source replica models (at a lower state of the art) and with no moderation emerge (see [Stable Diffusion](#)), the threat will become more real. In the case of the open-source Craiyon model (“DALL-E mini”), it is already possible to ask for images of public persons and to share AI-generated faces.



**A QUESTION OF RESOURCES AND INFRASTRUCTURE:** Replica models that are not subject to strict control mechanisms still lack enough data to train them adequately, as training these models from scratch is expensive and resource intensive. One expert interviewed mentioned that experiments with ruDALL-E, an open-source neural network that can generate images from texts and immediately translates the text prompt into Russian, for instance, showed that there were no data points in the model yet that would allow for presenting public figures in compromising positions. In less democratic countries, a lot of data is collected on citizens that can be easily used for the training of AI models. Two experts, therefore, voiced the concern that authoritarian state actors would eventually have the resources, infrastructures and vast amounts of data necessary to train such models in a smaller amount of time and could, perhaps, train them for targeting their opposition, which could then be used to put power in the hands of almost anyone.



**A QUESTION OF DETECTION MECHANISMS:** Experts outlined that the biometric verification systems that are currently in place do not always detect images of people generated from text prompts. This is already worrisome. Even if current forensic tools can relatively accurately detect an image generated by a text-to-image conversion tool, experts are afraid that content will keep on increasing to

a point where forensics cannot keep up with the sheer volume of production. We know that advances in detection mechanisms will always increase disinformation sophistication, with the synthetics becoming more ambitious by default,<sup>22</sup> fixing glitches that are currently detectable. If models do not adhere to self-regulation and authentication mechanisms, flaws in detection can and will fuel propagation.



**A QUESTION OF INOCULATION:** From a societal point of view, the threat level inherent in text-to-image generation depends significantly on the degree of society-wide resilience. One media expert said that if people refuse to approach social media with the same scepticism they would a tabloid newspaper, for example, this will give those spreading disinformation the ability to sway public opinion around information sources online, resulting in a reduced level of trust in news on social media.

## The Impact of Text-to-Image on Disinformation: Four Likely Scenarios



“If you can create photo-realistic images of politicians in incriminating situations, that probably, for a typical Internet user, carries more evidentiary power than a written statement that they did something politically damaging.”

*Micah Musser, Research Analyst at Center for Security and Emerging Technology*



### Scenario A: Falsified Evidence for Operations' Claims

The models' understanding of language allows for more flexibility and tractability in composing new images from natural language, which could have important applications for disinformation operations, namely, the creation of AI-produced imagery in direct response to a given false narrative, allowing for immediate fake evidence. Once the campaign messages are clear, images can be created as needed to support a false narrative. Even if most content policies do not yet allow for the retrieval of well-known faces from the training data, once the models become open access, it is still possible that such images can be introduced, which might then be used to generate harmful content misrepresenting public figures.

### Scenario B: Concealing the Inconvenient Truth

When synthetic content serves as fake evidence for an information operation's claims, the underlying strategy can be to generate a relatively high volume of synthetic images

<sup>22</sup> For more information on future disinformation threats, see Jan Beyer & Lena-Maria Böswald, [“On the Radar: Mapping the Tools, Tactics and Narratives of Tomorrow's Disinformation Environment”](#), Democracy Reporting International, June 2022.

to conceal “the true signal”, the fact that is considered to be unfavourable. Synthetic content could, therefore, undermine confidence in journalism and trustworthy sources of information. With a rise of highly believable synthetic content, even accurate content can be labelled as “false”, reducing trust in credible news institutions by sowing mistrust in legitimate information (and providers). There will be a lot of pressure on fact-checking and human rights organisations to prove the falsifiability of content. Disinformation actors will have the opportunity to use plausible deniability by declaring content to be synthetically generated, giving disinformation actors the upper hand, as fact-checkers face a higher burden of proof.

### Scenario C: To Label or Not to Label Content?

The challenges with deciding to label AI-generated content in the future can also have an impact on people’s trust in the information they are exposed to online. Two scenarios have already been addressed by OpenAI: the “implied truth” and the “tainted truth” effects.<sup>23</sup> While the former fosters people’s belief in unlabelled content as automatically true, because there is no warning label visible, the latter implies that any form of identifying and labelling false content will result in people doubting any information they come across, which could ultimately contribute to political apathy.

### Scenario D: Smaller Platforms – Growth over Security?

Disinformation operators could target newer platforms to test new or refined models and to fine-tune their efforts, as engaging content can go viral on smaller platforms quickly, and there are often only limited verification systems in place. While larger platforms deploy AI-enabled tools and human threat investigators to detect behaviour that violates their policies, smaller platforms often lack the requisite tools or human resource capacity – or simply the will – to apply mechanisms to detect harmful content. When growth is favoured over security, this creates fertile soil for the dissemination of entirely AI-generated content.

## **Who Are the Targets? Who Are the Perpetrators?**

### **Targets**

It is important to understand which groups are most vulnerable to suffering the consequences of the use of synthetic media produced by text-to-image generators. Who the most likely victims will be will depend on the context. The report, therefore, presents only some examples of risk groups brought by the experts we interviewed.

---

<sup>23</sup> Pamela Mishkin & Lama Ahmad, “[DALL·E 2 Preview – Risks and Limitations](#)”, GitHub, 24 May 2022.

## Vulnerable Groups and Minorities

Most interviewees cited minorities and other vulnerable groups as among the most likely main targets of media manipulation and synthetic content. One of the first groups most impacted by deepfakes were women, where their images were used to create pornographic content.<sup>24</sup> This type of content is manufactured without the consent and knowledge of the affected individuals. According to a report by DeepTrace, 96 per cent of online deepfake videos include pornographic content.<sup>25</sup>

One expert interviewed highlighted that text-to-image generation poses a particular threat to marginalised communities, in that they are those who suffer most from societal biases. Because some of the terms used in text prompts have historically been gendered, racialised and age-based, the models are being trained with datasets that reinforce these societal biases.

## Politicians / High-Profile Actors

Politicians and other high-profile actors are possible targets of the use of synthetic media, due to their relevance in public debates and access to sensitive security and financial information. Cybercriminals, for example, may attempt to impersonate government officials or CEOs to gain information or access financial resources from government institutions or private sector companies. Even though most text-to-image generators have output filters that do not allow for the portrayal of such actors, those that are open-access and, therefore, meet fewer regulatory standards (Craiyon or DALL-E mini) allow people to generate images of high-profile actors. Using unverifiable visual content to impersonate officials has the potential even to spark conflict.



“What happens if there is now a picture or video of Vladimir Putin shooting a civilian or an audio of Joe Biden saying something offensive? It is just a matter of time for this to happen, and partly is due to the democratisation of access to sophisticated technology to create fake images.”

*Hany Farid, Professor of Computer Science, University of Berkeley, California*



One key aspect when thinking of politicians as targets of synthetic media is the protection network they have around them. The level of popularity and power of a politician impacts directly how well they are protected from online threats. The protection network around the President of the United States, for example, probably has greater human resources capacity to verify and debunk manipulated content compared to that of a mayor or governor. Having more people in a team charged with

<sup>24</sup> For examples, see Madeline Brady, [“Deepfakes: A New Disinformation Threat?”](#), *op. cit.*, note 1.

<sup>25</sup> Giorgio Patrini, [“Mapping the Deepfake Landscape · Giorgio Patrini”](#) blog post.

tackling online threats will lead to a higher level of safety against the abuse of synthetic. Conversely, if we consider small-town politicians or ordinary members of a parliament, there might not be any real protection at all. Anecdotally, there are cases of local politicians who have quit their jobs because of this type of online harassment.<sup>26</sup>

### Journalists/Human-Right Activists

Two other groups that have weaker protection networks around them, and so are more likely to suffer from synthetic media abuse, are journalists and civil society activists working on political issues.



“People who are threatened by attacks on their credibility and by misinformation, and already face lack of support are generally more vulnerable than others. This is true for civil society, for people involved in electoral processes, for journalists and human rights defenders, especially in racist terms. If they are presented with manipulated content created out of thin air or fully synthetic media, they have far fewer defences, especially when they don’t have the public authority or the technical means to challenge and disprove them.”

*Sam Gregory, Program Director at WITNESS*



These groups are people who are targeted by state or corporate actors with a vested interest in doing so, and who don’t have access to effective ways to disprove manipulated content.

### Perpetrators



“Technology is not necessarily a threat in itself. Synthetic media are tools, like most technology, that can be misused, perhaps particularly easily. And that is, I think, the danger – the people who create this content with an agenda, who try to push malicious content.”

*Philipp Lorenz-Spreen, Research Scientist at Max-Planck-Institute*



When talking about perpetrators, we mean malicious actors that create harmful synthetic media with the intention of manipulating society and/or targeting specific groups through intentional misrepresentation. Perpetrators are context-specific, and they range from individuals to organised groups, including:

---

<sup>26</sup> Patrick Blisson, [“Why Local Officials Are Facing Growing Harassment and Threats”](#), Bloomberg.com, 29 June 2022.

## Authoritarian Regimes

Authoritarian regimes may attempt to create disinformation campaigns to push a narrative in their favour and convince citizens to believe the content they see online. Also, if citizens are unable to know with certainty which news content is true or false, this may overwhelm their critical thinking skills, leading to an inability to make informed political decisions.

## Cybercriminals

Synthetic media can take online criminality to the next level. Cybercriminals can use synthetic content to blackmail people into doing something, or to impersonate high-level actors to gain information or access to financial resources, for example. Text-to-image models are steadily becoming more advanced and able to reproduce more realistic, high-quality images, and access to these models is growing. Hence, it is increasingly likely these models are will be used for social engineering, which is a tactic used in many cybercrimes that manipulate human behaviour by toying with our trust, creating a sense of urgency, or playing on our sense of shame.

Cybercriminals are not necessarily from large groups or organisations. They can simply be individuals that have access to machine learning models and are able to create images generated by text prompts. They may create content to prank people, to falsely target public and private individuals, or to defame or to blackmail people.

There are many potential contexts in which malicious actors use synthetic media in order to gain some sort of advantage or to drive a narrative. One expert interviewed suggested that, independent of the context, perpetrators rely on trust networks. Trust networks refer to how people perceive the medium from which they consume information. If someone always reads a specific newspaper, for example, and trusts that the articles published there are fact-checked and based on actual events, that means they trust the information they are receiving. Perpetrators rely on people trusting different sources of information in getting them to believe the malicious content they are spreading. This scenario becomes dangerous in contexts where the majority of the population consumes information on social media, such as in Brazil, for example, where 79 per cent of the population uses WhatsApp as a news source.<sup>27</sup> There is no fact-checking or strategy to combat disinformation on the app, which makes the platform a great network for malicious actors to use to spread synthetic content, with the expectation that many – if not most – users will simply believe the “information” received.

---

<sup>27</sup> Jonas Valente, [“WhatsApp é principal fonte de informação do brasileiro, diz pesquisa”](#), Agência Brasil, 10 December 2019.

## 6. HOW PREPARED ARE WE?

### Are Our Societies Prepared?

Understanding how well our societies are prepared, and especially those that are less familiar with disinformation narratives and what synthetic media can do, is a key aspect in order to develop initiatives and efforts to raise awareness about and combat the threats posed by text-to-image generation.

Societal preparedness depends heavily on one crucial factor – digital literacy. The experts interviewed for this report suggested that societies that have a higher proportion of digitally educated people are better prepared to tackle the threats posed by text-to-image generation. This does not, however, automatically shield people from consuming disinformation in image format; it only suggests that fewer people tend to believe every piece of information they find online. A study by Harvard University revealed that, indeed, digital literacy is an important predictor of the ability to tell truth from falsehood when judging online content. However, although digital literacy enables people to better detect false content, it is not a guarantee that individuals will not share such content themselves.<sup>28</sup>

In this sense, digital literacy also influences who is more at risk of becoming a target, as digital skills decrease significantly with age and increase with a higher level of education,<sup>29</sup> making two groups the most vulnerable – elderly people and those with lower levels of education. Being able to understand how platforms work, as well as how to verify, classify, and even recognise manipulated content is increasingly important with the advances in AI. According to one expert interviewed, we need to

<sup>28</sup> Nathaniel Sirlin, Ziv Epstein, Antonio A. Arechar & David G. Rand. [“Digital Literacy Is Associated with More Discerning Accuracy Judgments but Not Sharing Intentions”](#), *Harvard Kennedy School Misinformation Review*, 6 December 2021.

<sup>29</sup> Sabine Knoll, [“Digitale Medienkompetenz in Deutschland: Studie Liefert Beunruhigende Ergebnisse”](#), *Der Spiegel*, 23 March 2021.

invest much more in our societies' digital skills, in the awareness of our populations, in building our own resilience to the impact of this type of content, and in learning how to navigate this complex new reality.

One expert highlighted the duality of societies being well prepared with regard to some factors, and not prepared at all with regard to others. On one hand, we are well prepared because in many ways we already understand how the technology works, whether it's from playing video games and seeing very realistic avatars, playing on a synthesiser app that reproduces the sound of a piano that is not real, or watching movies and seeing "realities" that do not actually exist. People know these are not real, they know these types of content are synthetically generated, and they can easily detect them. On the other hand, however, people remain very much susceptible to cheapfakes, because they cannot assess their sources or contexts.

“

“Society is very susceptible to shallowfakes [cheapfakes] because we generally struggle to understand how manipulation and framing works. We have a hard time evaluating sources and context properly. We do not always understand that in our daily media consumption there is already a lot of cutting, framing and editing of the content we see – with the tabloid press being a prime example.”

*Alexander Plaum, Innovation Manager at Deutsche Welle*

”

Regarding preparedness to deal with the most sophisticated technologies, one expert interviewed argued that we are not at all prepared for the development of synthetic media, specifically text-to-image capabilities. Only a limited number of people can grasp the potential of text-to-image conversion technology, as many of the text-to-image generators are not yet available to the public. Yet, even in cases where we can identify an image as AI-generated, the way we process information can still be a problem. While we may know rationally that the image is a fake, it still gets stored in our memory, and may subconsciously affect our outlooks and beliefs. This idea, labelled the “sleeper effect”, is worrying, because it shows that even people who are informed cannot fully defend themselves against the effects of the misuse of this technology.

“

“I think the way people epistemologically process information is not straightforward. Even a text without a source and no substantiating material does have a psychological impact on people. The mere exposure effect is real. And that is why the difference between a fake text supported by fake imagery and fake text only may not be as large as we would expect.”

*Micah Musser, Research Analyst at Center for Security and Emerging Technology*

”

Another expert, however, suggested that societies are less prepared because people use social media as their source of news and information. They highlighted that social media platforms were initially designed for entertainment and to connect people with family and friends, and not as places where they received information about the world. This is important, as people find it more difficult to distinguish between accurate information and disinformation on social media, such as on Facebook. A study shows that in Germany, a country where 87 per cent of the population has sufficient digital skills to conduct research online,<sup>30</sup> more than one-quarter of people surveyed believed that the number of likes for and comments on a post are a helpful indicator of that message's trustworthiness.<sup>31</sup> This demonstrates that many people are not fully aware of what the features of social media platforms imply, and how to infer whether information is true or not.

“

“If we think about solutions, no solutions can work without the people. We can do content moderation, we can hunt disinformation with fact-checkers, but the more sustainable approach, I think, is to give people the tools and context at hand to detect it themselves.”

*Philipp Lorenz-Spreen, Research Scientist at Max-Plank-Institute*

”

## Are We Technically Prepared for the Development of These Models?

With the advent of manipulated media online, we have seen advances in detection mechanisms for synthetic media. One of the most prominent of these is provenance technology, used to trace the origin of manipulated media through its metadata, or used as part of the product itself, such as visible digital watermarks.<sup>32</sup> This shifts the task of verification from the receiver to the producer of the visual information. The basic premise is that people need a way to know that the content they are seeing has actually been produced by the alleged source. The Coalition for Content Provenance and Authenticity (C2PA) has unified the efforts of the Adobe-led [Content Authenticity Initiative](#), Microsoft's [Project Origin](#) and the BBC to build an open technical standard for authentication.

<sup>30</sup> Initiative D21, [“Studie: Digital Skills Gap”](#).

<sup>31</sup> Anna-Katharina Meßmer, Alexander Sänglerlaub & Leonie Schulz. [“Quelle: Internet? Digitale Nachrichten- und Informationskompetenzen der deutschen Bevölkerung im Test”](#), 12 July 2022.

<sup>32</sup> For an overview of provenance technology initiatives, see Madeline Brady and Rafael Goldzweig, [“Deepfakes: How Prepared Are We?”](#), *op. cit.*, note 2.

Provenance technology can come in direct and indirect disclosure forms:

1. A digital watermark attached to media content to indicate whether the content has been manipulated;
2. A cryptographic meta-data signature, created at the moment of image creation, that authenticates the pixels that will be imbedded in the picture; or
3. An immutable hashing database that will maintain a record of where and when the image was taken.

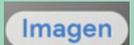
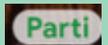
## Is Authentication at the Point of Creation Still Applicable for Text-To-Image Generation?

Forensic experts interviewed said that they believe provenance technology based on a metadata signature is fairly media agnostic and easily implementable into text-to-image technology, if so desired by the service providers. They even said that watermarking or hashing databases, to facilitate downstream identification, should be required of AI companies. They all said they doubted that a visible digital watermark, as implemented by all AI providers, could serve as the sole authentication safeguard once these more powerful models become freely accessible, as this can easily be cut off or edited with Photoshop. In this case, they said, using such watermarks is more of a branding technique than a protection mechanism.

One possibility would be to record (through blockchain hashing) each image created, and maintain a general database where these can be looked up. Depending on the production speed and the amount of images produced, however, this could pose a scaling problem. Another possibility would be to introduce an invisible digital watermark directly into the image, which would be difficult to remove. In this case, no storage is required. It is slightly less secure than hashing, but can handle such a large number of images very easily.

While the idea of provenance technology is complex to implement systematically because of its reliance on blockchain technology and the fact that platforms often remove metadata from media content, visible digital watermarking is already used by most of the models explored in this paper:

Table 2. Responsible AI practices followed by the AI service providers by model

Model	Developer	Open access	Controlled access <sup>33</sup>	Visible digital watermarks	Filtered input <sup>34</sup>	Filtered output <sup>35</sup>
<b>DALL-E 2</b>	Open AI	No	Yes, but eligible users can soon acquire full usage rights to commercialise the images. <sup>36</sup>	Yes 	Yes	Yes
<b>Craiyon</b>	Boris Dayma <sup>37</sup>	Yes	No	No	No. The model was trained on unfiltered web-scale data, limited to pictures with descriptions in English.  Initial testing demonstrates that the model may generate images that contain negative stereotypes of minorities.	Yes
<b>Imagen</b>	Google	No	Yes	Yes 	The model relies on text encoders trained on uncurated web-scale data, thus inheriting the social biases and limitations of large language models.  For model training, it used the LAION-400M dataset, which is known to contain a wide range of inappropriate content, including pornographic imagery, racist slurs and harmful social stereotypes.	A subset of its training data was filtered to remove noise and undesirable content.
<b>Parti</b>	Google	No	Yes	Yes 	The model is trained on large, mostly uncurated datasets, obtained from the web with little oversight.  The primary training data is selected and highly filtered, to minimise the presence of harmful and unsafe content. The model is to be recalibrated in the near future.	Yes
<b>GLIDE</b>	Open AI	Released a smaller diffusion model and noised CLIP model on <a href="#">GitHub</a>	Yes	No	The model is trained on large, often noisy, image-text datasets that are known to contain biases regarding people of different backgrounds.  The model filtered out training images containing people, to reduce the capabilities of the model in many cases of people-centric problematic use, as well as violent images and hate symbols.	No, but the data filter had a sufficiently low false-negative rate.

<sup>33</sup> The model is only available to a selected group of signed-up researchers to guard against misuse.

<sup>34</sup> The model's input data is controlled or adapted to mitigate the use of graphic/explicit training data.

<sup>35</sup> The model includes output filters that prevent users from generating what it deems to be inappropriate images.

<sup>36</sup> For more information, see "[DALL-E Now Available in Beta](#)", OpenAI, 20 July 2022.

<sup>37</sup> The first version of Craiyon (DALL-E Mini), inspired by Open AI's DALL-E, was produced at a hackathon organised by Hugging Face and Google in July 2021.

## Is There a Way to Dissect the Level of Noise in a Picture to Identify whether a Photo is Real or AI-Generated?

DALL-E 2, Imagen and GLIDE are all based on diffusion models that destroy training data by adding noise to the picture, and then learning to recover the data by reversing the noising process. As a consequence, it is important to know whether it is forensically possible to dissect the level of noise to differentiate between a real and a generated image.

Both forensic experts interviewed for this report said that images leave more statistically detectable traces than simple textual information. As a consequence, they say, it is technically possible to distinguish the type of pixel-level noise artefact in a picture. If not done already, DALL-E 3, for instance, could be very easily trained to add imperceptible amounts of noise to any specific image to make it easier to detect. The question remains, however, of how sustainable or vulnerable to counterattack this technique is. Another forensic technique – lighting analysis – that estimates the 3D lighting environment of an object in a picture may prove even more useful in distinguishing synthesised from real images in the future.<sup>38</sup>

### Are the Platforms Prepared?

Many social media platforms have already adopted policies to better prepare themselves to deal with manipulated media content and properly address the threat and misuse of synthetic media for disinformation purposes. An overview of current policies in practise can be found below.

Nonetheless, no platform has yet tackled the specific issue of merging multiple tools or addressing the risk AI-generated image based on text prompts pose. This is mainly because, from a platform perspective, those images alone – without fake text as supporting evidence – would not pose a different threat than misleading synthetic or manipulated media content; they would simply fall under the same manipulated media guidelines and standards. However, some platforms (see Meta’s work on “radioactive data”<sup>39</sup>) are investing in prevention mechanisms that can support provenance technology. For such an authentication mechanism to be successful, platforms would need to be highly involved in the further development and potential implementation of such technology. Other research focuses on AI-powered tools to address and help detect misinformation, trained on Wikipedia data.<sup>40</sup>

<sup>38</sup> For more information on lighting analysis, see Hany Farid, “[Lighting \(In\)Consistency of Paint by Text](#)”, University of California, Berkeley, 30 July 2022.

<sup>39</sup> Hervé Jégou, Matthijs Douze & Alexandre Sablayrolles, “[Using ‘radioactive data’ to detect if a dataset was used for training](#)”, Meta AI, 5 February 2020.

<sup>40</sup> “[How AI could help make Wikipedia entries more accurate](#)”, Tech at Meta, 11 July 2022.

Platform (not exhaustive)	Policy	Criterion	Response																				
<b>Twitter</b> 	<a href="#">Synthetic and Manipulated Media Policy</a>	The policy forbids users from sharing altered content that may confuse people or lead to harm; in some cases, Twitter may label tweets as containing misleading media to provide users with more context.	<p><b>If applicable, content is labelled and visibility is reduced, or content is removed.</b></p> <table border="1"> <thead> <tr> <th>Is the media significantly and deceptively altered or fabricated?</th> <th>Is the media shared in a deceptive manner?</th> <th>Is the content likely to impact public safety or cause serious harm?</th> <th></th> </tr> </thead> <tbody> <tr> <td>✓</td> <td>✗</td> <td>✗</td> <td>Content may be labeled</td> </tr> <tr> <td>✓</td> <td>✗</td> <td>✓</td> <td>Content is likely to be labeled, or may be removed.</td> </tr> <tr> <td>✓</td> <td>✓</td> <td>✗</td> <td>Content is likely to be labeled.</td> </tr> <tr> <td>✓</td> <td>✓</td> <td>✓</td> <td>Content is very likely to be removed.</td> </tr> </tbody> </table>	Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?		✓	✗	✗	Content may be labeled	✓	✗	✓	Content is likely to be labeled, or may be removed.	✓	✓	✗	Content is likely to be labeled.	✓	✓	✓	Content is very likely to be removed.
Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?																					
✓	✗	✗	Content may be labeled																				
✓	✗	✓	Content is likely to be labeled, or may be removed.																				
✓	✓	✗	Content is likely to be labeled.																				
✓	✓	✓	Content is very likely to be removed.																				
<b>Meta</b> 	<a href="#">Community Standards: Manipulated Media</a>	<p>The policy forbids videos that have been edited or synthesised, beyond adjustments for clarity or quality, in ways that are not apparent to an average person and would likely mislead an average person to believe the contents are authentic.</p> <p>The item is the product of AI or ML, including DL techniques, that merges, replaces or superimposes content onto a video, making it appear to be authentic.</p>	<p><b>If applicable, content is removed.</b></p> <p>Manipulated videos that do not meet this standard are generally eligible for fact-checking and receive a specific rating for “altered” content.</p>																				
<b>Meta</b> 	<a href="#">Community Standards: Misinformation</a>	<p>The policy prohibits content that is likely to directly contribute to the risk of imminent physical harm, interference with the functioning of political processes, and certain highly deceptive manipulated media.</p> <p>It further prohibits content and behaviour in areas that often overlap with the spread of misinformation: <a href="#">fake accounts</a>, <a href="#">fraud</a> and <a href="#">coordinated inauthentic behaviour</a>.</p>	<p><b>If applicable, content is removed. For all other misinformation, they partner with third-party fact-checking organisations to review and rate the accuracy of the most viral content.</b></p>																				
<b>YouTube</b> 	<a href="#">Misinformation Policies: Manipulated Content</a>	The policy includes content that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context), and that may pose a serious risk of egregious harm.	<b>If applicable, content is removed.</b>																				
<b>TikTok</b> 	<a href="#">TikTok Community Guidelines: Integrity and Authenticity</a>	The guidelines prohibit synthetic or manipulated content that misleads users by distorting the truth of events in a way that could cause harm.	<b>If applicable, content is removed, accounts are banned, and it is more difficult to find harmful content in recommendations and search.</b>																				
<b>Reddit</b> 	<a href="#">Reddit Policy on Impersonation</a>	The policy forbids content that impersonates individuals or entities in a misleading or deceptive manner. This also encompasses domains that mimic others, as well as deepfakes or other manipulated content presented to mislead, or be falsely attributed to an individual or entity.	<b>If applicable, content can be reported and deleted (a profile ban).</b>																				

# 7. WHAT IS CURRENTLY BEING DONE?

## Policymakers

This section will focus on the efforts of policymakers in the European Union to better regulate the online sphere and, consequently, tackle the threats posed by synthetic media. It focuses on the EU, as current regulatory trends there are pioneering in this field, and could serve as examples for countries outside of the bloc in their own legislation, leading to a global impact. Currently, there are two important legislative tracks being followed:

- 1) **The Digital Services Act (DSA):** The DSA aims to create a safer digital space in which the fundamental rights of all users of digital services are protected, and to establish a level playing field to foster innovation, growth and competitiveness, both in the European Single Market and globally. The rules specified in the DSA primarily concern online intermediaries and platforms, such as online marketplaces, social networks, content-sharing platforms, app stores, and online travel and accommodation platforms.<sup>41</sup>
- 2) **The Artificial Intelligence Act:** The AI Act is a proposed European law on artificial intelligence, and would be the first law on AI adopted by a major regulator anywhere. The law assigns applications of AI to three risk categories. The first covers applications and systems that create an unacceptable risk, such as government-run social scoring of the type used in China, which would be banned. The second covers high-risk applications, such as a CV-scanning tool that ranks job applicants, which would be subject to specific legal requirements. The third covers applications not explicitly banned or listed as high-risk, which would largely be left unregulated.<sup>42</sup> The EU AI Act could become

<sup>41</sup> European Commission, [“The Digital Services Act Package | Shaping Europe’s Digital Future”](#).

<sup>42</sup> European Union, [“The Artificial Intelligence Act”](#), 7 September 2021.

a global standard, determining to what extent AI has a positive rather than negative effect on people's lives, wherever they may be. The EU's AI regulation is already making waves internationally. In late September, Brazil's Congress passed a bill that creates a legal framework for artificial intelligence.<sup>43</sup>

Legislators are working on these regulations because they understand the necessity to create standards and enforcement mechanisms for AI models and novel technologies. The AI Act, one regulatory expert interviewee said, was not easy to develop. It is the first comprehensive proposal that applies to all sectors and that addresses diverse kinds of risks in various applications. It is important, however, to create strong enforcement mechanisms and binding obligations for platforms to ensure a healthier digital environment.



“The AI Act draft seeks to strike a balance between innovation and risk mitigation to harness AI technologies that can be trusted, add value, and are used for our common good while addressing risks without hindering innovation.”

*Yordanka Ivanova, Legal and Policy Officer for the European Commission*



One important aspect of the AI Act is how the term “AI” was defined. According to one expert, it considers AI as a family of novel technologies and, even though there is no mention of specific models, they argued that the definition should also cover machine-generated content. The broad definition was meant to safeguard against restrictions that could hinder innovation. Policymakers want to enable innovation using AI technologies that can be trusted and used for our common good. This is also the reason only high-risk practices would be prohibited.

When it comes to text-to-image generation, there is still no mention in the proposed laws of content generated this way. The expert suggested that they have not yet addressed this issue because it is a new technology, and they need more studies and evidence before they can categorise it in their risk assessment. Once again, the broader definition of high-risk categories appears to have been used so that regulators can add more cases and adapt the law to technological and market developments.

## CSOs

CSOs work on a range of different topics to ensure that the risks and threats posed by text-to-image generators are minimised:

<sup>43</sup> Portal da Câmara dos Deputados, [“Câmara aprova projeto que regulamenta uso da inteligência artificial – Notícias”](#), 29 September 2021.

## Advocacy for Legislation

With the likelihood that the Digital Services Act (DSA) and AI Act will become law, CSOs like [EPD](#) and [Article 19](#) are advocating for improvements in legislation to properly address manipulated media content.

Regarding the DSA, most of the advocacy efforts are pushing for greater transparency when it comes to the AI used at various stages of ranking decisions, such as the choice to amplify disruptive content over high-quality content. This transparency is necessary in order to understand how platforms' internal policies rank and categorise different types of content in order to take any necessary action against them.

CSOs are also pushing for more specific classification and assessment in the DSA of high-risk AI practices. Articles 26 and 27 of the DSA require very large online platforms to carry out risk assessments at least once a year, to identify systemic risks of the dissemination of illegal content, of any negative effects on the exercise of certain fundamental rights, and of the intentional manipulation of their service. In addition, these platforms are required to take reasonable, proportionate and effective measures to mitigate these risks, under the supervision of the European Commission, in cooperation with the European Board for Digital Services and the Digital Services Coordinators. CSOs argue that these articles are too vague and fail to meet the legality test under international human rights law, i.e., they are insufficiently precise to enable platforms, users and others to foresee how any risks to human rights will be addressed.

The concerns related to the AI Act follow the same logic, focusing on it using vague definitions and concepts. Most CSOs believe that more concrete criteria are needed for a future ban on more forms of harmful AI. One example is the risk-based approach, by which the AI Act states that some technologies constitute "unacceptable risks". The CSO [Article 19](#) argues that this definition is flawed, because the harm of an unacceptable risk only incorporates physical or psychological individual harm, and ignores the fact that AI systems often impact communities in intangible ways that are difficult to prove or predict. Others call for more forms of AI to be given a high-risk label.

## Media Literacy and Detection Techniques

CSOs play an important role in empowering people to understand and learn more about how artificial intelligence works. There are many organisations that work on giving people access to information and knowledge about advances in AI and tools to detect manipulated content online. Media literacy can counter disinformation while protecting freedom of expression, and it is important that more people have digital skills if we want to have a healthier and more democratic online space.

Many CSOs are working on teaching more practical detection techniques, for example, teaching people how to spot manipulated images online. For example, [First Draft](#), an

organisation that provides practical and ethical guidance on how to find, verify and publish content sourced from the social web, created the SHEEP approach.<sup>44</sup> [SHEEP](#), which stands for source, history, evidence, emotion and pictures, is a checklist people can use to help them make sure they can believe what they are seeing. Another example is the [Learn to Discern \(L2D\)](#) media literacy training provided by the [International Research & Exchanges Board \(IREX\)](#). IREX has developed a media literacy curriculum that is taught in classrooms, libraries and community centres in Ukraine, helping people to develop healthy habits for engaging with information online.

## Fact-Checking

Fact-checking organisations are key actors in verifying information that circulates online. In 2021, more than 100 fact-checking organisations signed a [Code of Principles](#) to establish standards for their commitment to principles including transparency, non-partisanship and fairness.<sup>45</sup>

Fact-checkers are extremely important with regard to synthetic media, and especially to images that are used in disinformation campaigns. It is important that fact-checkers understand how AI-generated media that is based on text prompts works, in order to detect it properly. They are also an excellent source people can rely on to verify news and content when facing disinformation campaigns. One example is the [European Digital Media Observatory \(EDMO\)](#) initiative to build a comprehensive “map” of [initiatives](#) in the European Union that provide fact-checking. For example, Deutsche Welle has a [fact-checking repository in Germany](#), where they debunk, explain and use in-depth research techniques to separate fact from fiction.

## Code of Practices

Synthetic media pose a challenge, in that it is difficult to distinguish the content it produces from reality. To overcome this challenge, we need to establish best practices for methods, tools, journalistic inquiry and media literacy. Partnership for AI (PAI) has been a leading CSO, working since 2018 with actors around the world to develop a Code of Conduct for the use of synthetic media.

This [Synthetic Media Code of Conduct](#) will be developed with leadership from WITNESS, Adobe, Microsoft and others in the synthetic media ecosystem. The aim is to craft guidelines that will influence norms and behaviours across sectors for those developing synthetic media technologies, creating synthetic media and distributing

---

<sup>44</sup> First Draft, [“Think ‘Sheep’ before You Share to Avoid Getting Tricked by Online Misinformation”](#), 9 December 2019.

<sup>45</sup> IFCN, [“The Commitments of the Code of Principles”](#).

synthetic media.<sup>46</sup> Initiatives of this type are crucial, as preparing society and various actors for a positive synthetic future requires introspection and action. The Code of Conduct will answer technical questions related to creating and addressing synthetic media, and will also examine policies and social infrastructure that influence how this technology is being developed and used. This Code of Conduct is an important step, as it is a multi-stakeholder effort involving CSOs, technology companies and other actors, demonstrating that cooperation between all of these actors is key to ensuring that ethical practices are in place in the online sphere.

### Provenance Technology

As already discussed in this report, provenance technology is fundamental to the detection of synthetic media. The [Coalition for Content Provenance and Authenticity \(C2PA\)](#) addresses the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content.<sup>47</sup>

The Coalition unites the efforts of policymakers, academics and industry leaders, and focuses exclusively on the development of open, global technical standards to channel content provenance efforts. The coalition's work includes the development of best practices and reference designs for applying these technical standards; the promotion of global adoption of digital provenance techniques; and ensuring that content remains accessible, even with digital provenance techniques applied.

### Defence of Journalists and Human Rights Activists

Journalists and human rights activists are among the most vulnerable groups to existing problems of media manipulation, state violence, gender-based violence and misinformation and disinformation. [WITNESS](#) is an organisation that pushes for the solutions needed by journalists and human rights defenders worldwide. One of their projects focuses on the emerging potential malicious uses of AI-generated synthetic media, and how we can push back to defend evidence, the truth and freedom of expression from a global, human-rights-led perspective.

---

<sup>46</sup> Claire Leibowicz, "[PAI Developing Ethical Guidelines for Synthetic Media](#)", *Partnership on AI* (blog), 10 March 2022.

<sup>47</sup> The Coalition for Content Provenance and Authenticity, [c2pa.org](#).

## 8. WHAT SHOULD BE DONE NEXT?

The threat posed by text-to-image generation is not limited to one group of people, but is an overarching issue across different domains. As the threat assessment earlier in this report demonstrates, there is a strong likelihood that evolving technology can narrow the gap between truly authentic and fake, but relatively plausible content. The following calls for actions and policies touch upon text-to-image conversion technology and its innovative and “disinformative” potential, reflecting upon the experiences of the experts interviewed for this report and their requests for change in the field:

### Increasing Protection Mechanisms

1. **We call for a binding and standardised “AI responsibility” Code of Conduct for AI service providers that goes beyond self-regulation.**



Provenance technology empowers consumers to assess whether what they are seeing is trustworthy, and provides validation for unchanged content during distribution. It is only applicable, however, at the point of creation of the image. This puts the onus on the AI industry to foster this innovation. Service providers should, therefore, be obliged to maintain an industry-wide, open standard for authentication of content. An exchange of best practices with other model developers is encouraged in this regard.

2. **We call on code-hosting platforms to play their parts in the authentication process by enforcing “product safety” standards.**



Code hosting platforms (e.g., GitHub or Hugging Face) publish text-to-image generation source code online. There should be a binding check system built into the platforms to verify digital traces and the authenticity of content before distribution.

3. **We encourage AI companies developing text-to-image generators to think about corporate responsibility and work with filters to reduce the risk of biased models.**



Most text-to-image models are trained on unfiltered, large-scale data, limited to pictures with

descriptions in English and text encoders trained on uncurated data. While this approach enables rapid algorithmic developments, datasets have produced biased outputs. With advances and refinements in model development, it would be better to use filtered subsets or better classification models of such large-scale training data, so as to prevent biased and undesirable results. Another option would be the use of synthetic data to train the model from the outset.

**4. We suggest constant model evaluation and system cards for potential harms of text-to-image generation.**



Initial testing demonstrates that the models may generate images that contain negative stereotypes against minorities. It is key, therefore, to implement input data audits and comprehensive dataset documentation. AI companies should further develop evaluation metrics to inform a responsible and sustainable model release.

**5. We recommend the development of trusting and cooperative relationships between regulators and the AI industry to jointly create minimum standards.**



Legislative experts interviewed for this report suggested the implementation of model trials in a controlled environment, under the supervision of the regulator, not to stifle innovation but, rather, to address risks, to allow the AI companies and the regulator to learn together, and to find the appropriate mitigating measures. This could be executed by means of an “AI sandbox”<sup>48</sup> to implement provisions with the help of legislators before the AI Act is adopted for implementation in Europe. Considering the dynamics in the field, effective minimum standards can only be developed through cooperation between regulators and model developers to allow for innovation. Newer or less-self-regulated text-to-image generation models might not include these standards. Here, there should be strict rules dictating high penalties and the removal of offending content.

## Encouraging Platform Transparency

**6. We expect social media platforms to apply stricter community standards for a combination of synthetic media.**



As no platform yet has touched upon merging multiple tools and the threat text-to-image generation might pose, these should assess its risk potential more specifically, by developing safety standards that counter the spread of synthetically produced disinformation, but that do not flag synthetic content by default, thus not targeting content that is not harmful. This can

---

<sup>48</sup> The European Commission’s sandbox approach aims to bring competent authorities close to companies that develop AI to define best practices that will guide the implementation of the [Artificial Intelligence Act](#). This would also ensure that the legislation can be implemented in two years. The regulatory sandbox is a way to connect innovators and regulators and provide a controlled environment for them to cooperate.

be achieved by the standardisation of synthetic content-detection integration into mainstream platforms.

- 7. We encourage social media platforms to be more transparent about policy enforcement data.** It is still unclear how much synthetic media can be found on platforms, whether this embodies an imminent threat to online discourse, and how the policies that apply to synthetic and manipulated media are being enforced. Without access to enforcement data, it is difficult to assess the effectiveness of these policies and, thus, how much of a threat to public discourse text-to-image generation can pose.



## Going Beyond Technical Solutions

- 8. We recommend the introduction of media literacy programmes that apply innovative sensitisation formats.** Closing the knowledge gap for different communities is essential to reducing the risk of this technology being used to harm public discourse. The media experts interviewed for this report said that an ideal training module would consist of traditional elements of awareness-raising (how does the technology work?) with a combination of gamification elements (how can one detect manipulated media?) that can tackle gaps playfully.



- 9. We advise CSOs and media outlets to shift their focus from debunking to pre-bunking.** Fact-checking and debunking techniques to correct incorrect information are ineffective for people who already believe false content. Educating audiences to discern and be watchful for cheapfake content, and about the prospects of text-to-image generation being used to promote disinformation will be key as a form of inoculation against misinformation.



- 10. We propose an exchange forum idea for effective collaboration between researchers, CSOs and tech platforms.** Only a strong ecosystem that closely follows developments in synthetic media and informs both public debates and regulatory approaches can build societal resilience. So far, no proper coordination mechanism to support this is in place. Dialogue, in a forum format, between different stakeholders and on policy diversity should, therefore, be encouraged.



