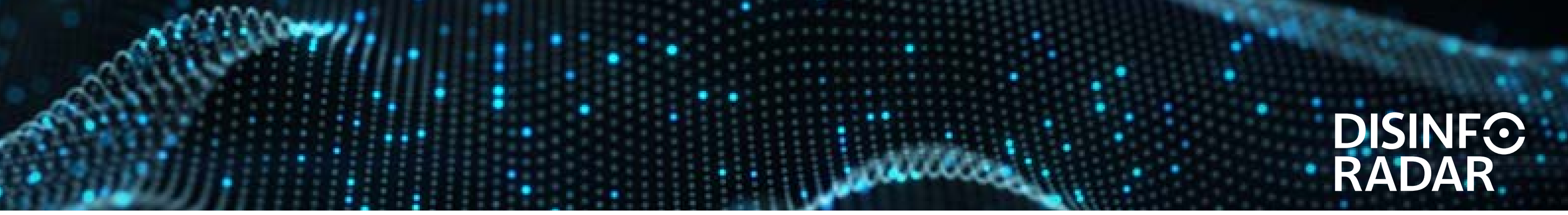


An Ambivalent Alliance: How Authoritarian Regimes Use – and Fear –Generative AI

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

Summary

- Authoritarian regimes historically have had a complex relationship with technology, leveraging it for control, yet fearing its disruptive power.
- The advent of generative AI introduces new challenges for authoritarian governments, which perceive these innovative tools as potential threats to their political authority.
- At the same time, authoritarian states will leverage generative AI for their own purposes, as the example of Kandinsky 2.1, the Russian text-to-image model, shows.
- The fact that generative AI models are now being developed in authoritarian states has the potential to make regulation in democratic states ineffective. The user-friendly nature of these models creates a proliferation risk – they may spread their intentional biases well beyond the borders of their countries of origin.
- In response, democratic states should prioritize the development of technologies capable of detecting outputs produced by such models, thereby enabling them to alert their citizens when they are exposed to such content.
- In light of the global concern regarding the authenticity of online content, the EU and the United States should discuss this issue with Chinese authorities and those in other states, in order to explore whether there is a possible minimum common denominator for cooperation.

Introduction

Throughout history, authoritarian regimes have had a paradoxical relationship with technology. On the one hand, they have been acutely aware that technological progress has the potential to upend established power structures and erode their authority. For example, several authoritarian states, including Cuba, North Korea, and others, **temporarily banned public ownership of mobile phones,** due to their potential for disseminating uncensored information and facilitating the organization of protests. Numerous authoritarian regimes have endeavoured to seal off (at a minimum) some segments of the internet, with China's **Great Firewall** being the most intricate system to date. This apparatus has morphed the internet into a walled garden, effectively barricading a myriad of global online services.

On the other hand, these regimes rely on technological innovations to exert control over their populations, quash dissent, and undermine democracy in other countries. By leveraging big data and machine learning to monitor and score individuals based on **social behaviour**, China is pioneering a new approach to extensive social control that illustrates how states can attempt to bring citizens in line. Russia has long used **troll armies, automated bots, and proxies** to influence social media as low-cost and potentially effective weapons of asymmetric warfare. States like Hungary have also used **sophisticated spyware** to control critical journalists.

As we enter the era of generative AI, with a myriad of tools emerging that allow the creation of synthetically generated text, audio, and even video, the ambivalence of autocratic regimes towards technology is again becoming clear. While much has been written about the potential for malicious actors, including authoritarian states, to use generative AI for disinformation, **propaganda, and foreign influence campaigns,** little attention has been paid to the disquiet among autocrats caused by new AI tools such as ChatGPT, Midjournay, and StableDifusion.

Case Study – China: On the Way to Regulation

The global debate on how to regulate generative AI has not left authoritarian regimes untouched. On the contrary, these regimes have responded with a heightened sense of alertness. There is a growing fear among them that generative AI tools, particularly those originating from the West, could threaten their authority and stability. For example, governments in Russia, China, Iran, and North Korea have found ways to **block access to tools like ChatGPT,** as they fear that such tools could be used to undermine their authority.

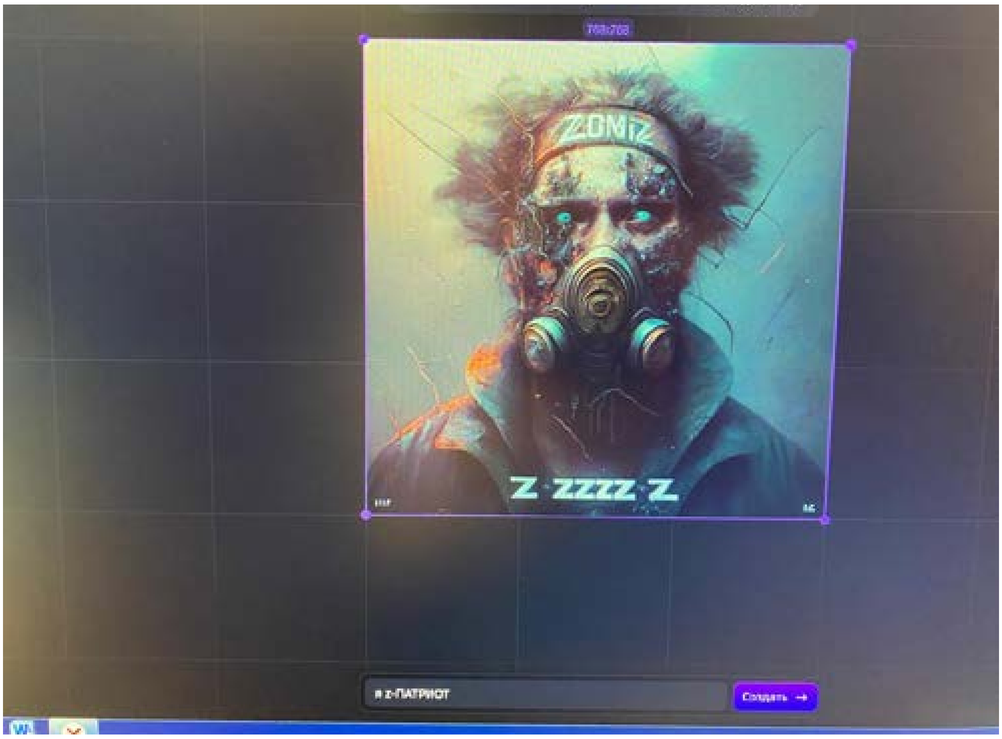
Yet the challenge that generative AI poses to autocratic policymakers goes beyond the role of tools like ChatGPT developed in the labs of Silicon Valley. China is itself a leading country in the development of AI, and has already seen the **widespread adoption of generative AI for commercial use.** It now faces domestic controversies and challenges related to privacy, security, and ethical and socioeconomic concerns. In response, the Cyberspace Administration of China (CAC) has released a draft policy, the **Administrative Measures for Generative Artificial Intelligence Services.** The proposed regulations reflect some concerns that democracies also face, such as those related to non-discrimination, bias, and the quality of training data. In fact, some of the requirements proposed by the bill, such as requiring **AI-generated content to be clearly labelled** and the establishment of complaint mechanisms, mirror those in the debates in democracies.

The proposed regulations in China's draft Administrative Measures for Generative Artificial Intelligence Services reflect not only the country's concerns about the **violation of privacy laws, however,** but also the tight nature of Chinese censorship. The bill would require companies to submit security assessments to authorities before launching their AI offerings for the public, and it explicitly aims to ensure that AI-generated content follows "**Socialist Core Values**" – a nebulous term that has for years been defined as the **general decorum** expected of Chinese citizens — and not to subvert state authority. Thus, as with the internet, the Chinese government appears to be trying to ensure that the proliferation of generative AI does not counteract its propaganda efforts.

Case Study – Russia: Image-Generation for Propaganda

Little noticed outside of Russia, the country had its own controversy about a generative AI tool, Kandinsky 2.1, which is a text-to-image generator introduced by Sberbank, Russia's largest bank. Examining the events surrounding Kandinsky 2.1 in greater detail sheds light on the evolving landscape of generative AI and how authoritarian regimes might assert greater control over and exploit the output generated in their respective countries.

Kandinsky 2.1 has recently drawn the attention of the Russian government. Initially praised by the tech community for its impressive speed in generating images, the model has since become the focus of controversy. On 26 April 2023, Sergey Mironov, a prominent Russian lawmaker, publicly complained about the online tool, alleging that it consistently generated negative images of the country. He went so far as to call the model “rogue” and “designed” to generate such negative images. Using Fusion Brain, a Kandinsky 2.1-powered online tool, Mironov attempted to create images using the letter Z, a symbol that has been associated with Russia’s military aggression against Ukraine. Yet the output resulted in dystopian or surreal imagery, rather than propagandistic narratives.



z-ПАТРИОТ

Image 1 — A screenshot taken from Sergey Mironov’s Twitter feed before modifications were made to the model

While the developers of Kandinsky 2.1 initially did not comment on such complaints, Twitter users have noticed a recent change in the text-to-image model. They pointed out that the model had started to produce an odd output when specific prompts were used. Prompted with “Z Patriot”, for example, it had stopped producing random imagery and, instead, repeatedly generated the same images of exaggeratedly masculine men with the Z symbol. Similarly, prompts like “Russian patriots” now generate images of cheering crowds holding the Russian flag, rather than generating random images. Given the probabilistic nature of text-to-image generation models, these outputs appear to be predetermined and, therefore, not actually generated by AI. In fact, it suggests that the developers of Kandinsky 2.1 may have bowed to state pressure and tweaked the system in such ways that it generates goodwill on the part of the authorities.

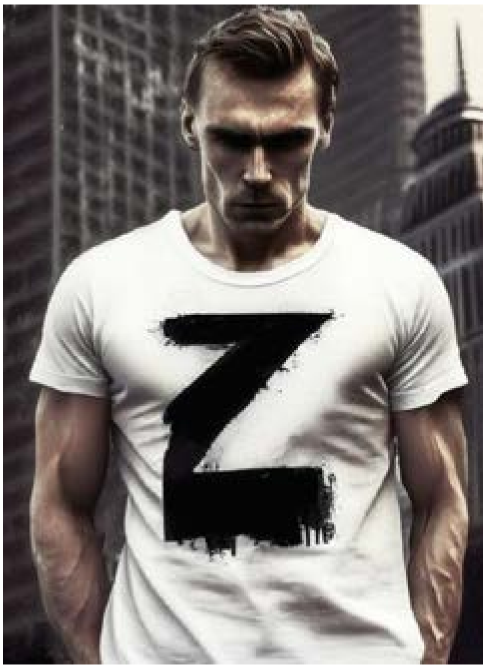


Image 2 — Output produced with Fusion Brain based on the prompt “Z Patriot” after modifications were made to the model.

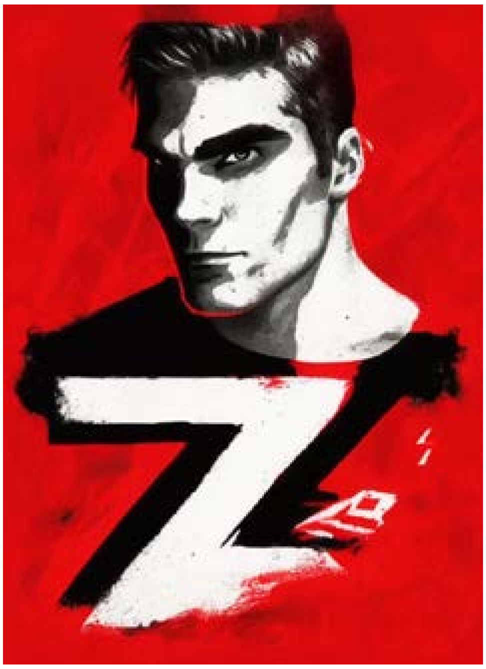


Image 3 — Output produced with Fusion Brain based on the prompt “Z Patriot” after modifications were made to the model.

The recent events surrounding the Russian state’s influence on Kandinsky 2.1 have highlighted the dangers of the interplay between authoritarian regimes and the emerging power of generative AI tools – the regimes can insist on certain design choices, ways of training models, or targeted tweaking to make them useful as propagandistic tools. While the modifications made by Sberbank, as the developers of Kandinsky 2.1, seem to have been rudimentary, it is possible that in the future we might see more powerful alliances between governments in authoritarian states and producers of generative AI models. Willingly or forced, developers might use explicitly biased training data, ensuring that the output produced is in line with state propaganda.

The issue goes far beyond war related propaganda, however. A DRI review of content produced by Kandinsky 2.1 has revealed how the model already facilitates the production of racist, antisemitic, or conspiracy-related content. Using prompts like “Jew”, “Racism”, or “Who controls the world” produces images that are well suited to hateful image campaigns.

The images below are sensitive and only shown here for research purposes.

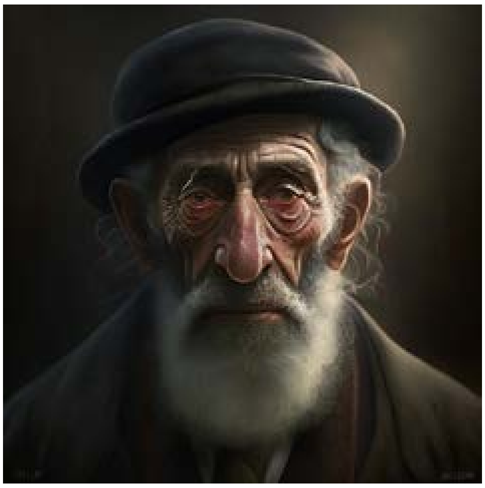


Image 4 — An image produced by Kandinsky 2.1 with the prompt “Jew”.

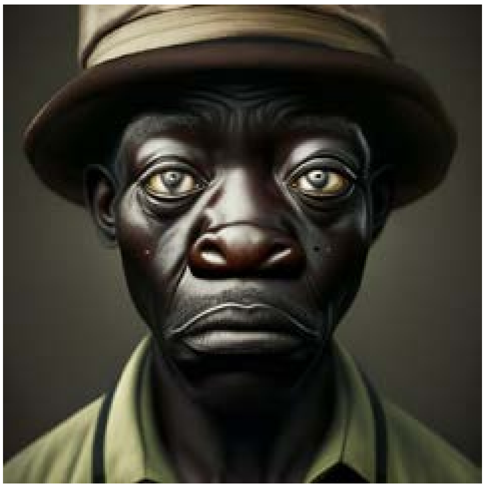


Image 5 — An image produced with Kandinsky 2.1 with the prompt “Racism”.

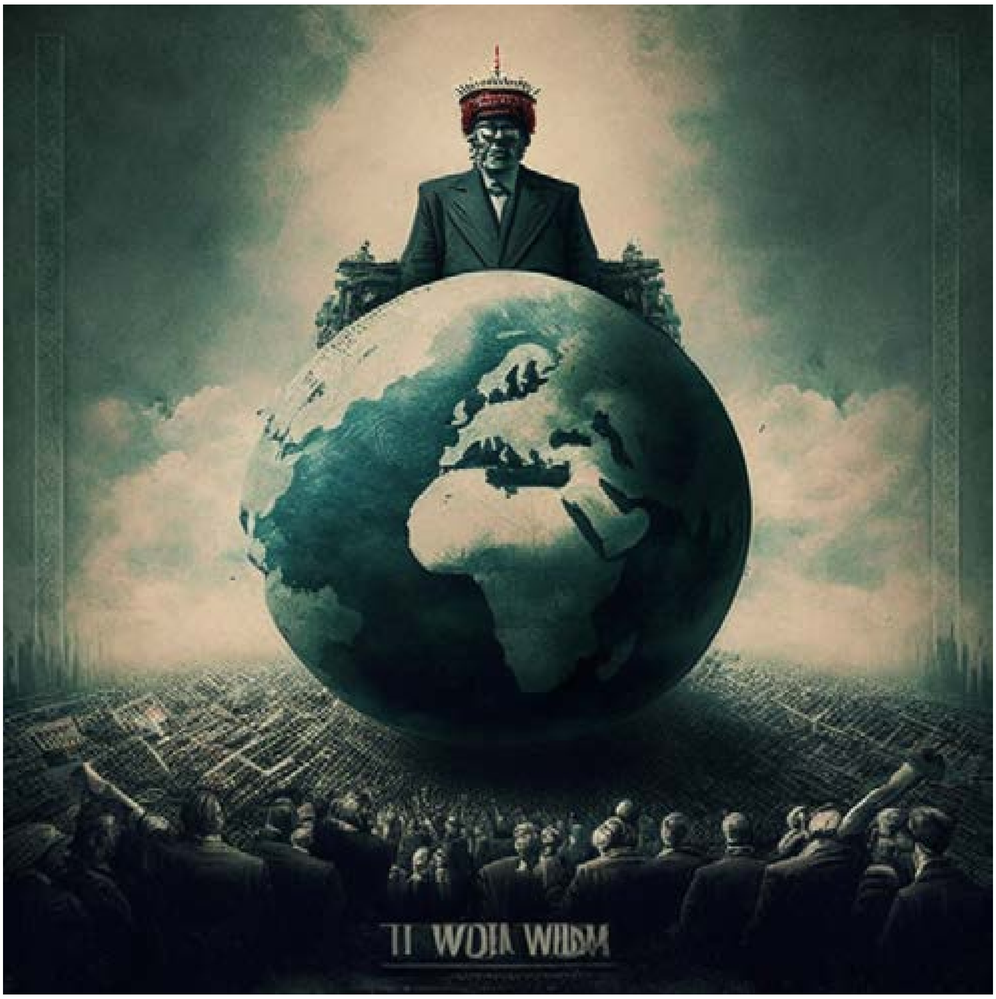


Image 7 — An image produced with Kandinsky 2.1 with the prompt “Who controls the world?”

The Need for Detection

Generative AI models developed in authoritarian countries – with possible state involvement – have implications that extend beyond the confines of these states. With user-friendly online tools powered by these models, they are becoming increasingly accessible globally. This ensures that the biases and propaganda originating from these models' home countries will proliferate far beyond their borders.

Attempting to regulate models developed in states ruled by authoritarian regimes is a difficult task, as they will remain beyond the reach of authorities in democratic states. On the European level, for example, the AI-Act, which is currently being drafted, encompasses a broad scope of application (including outputs "used in the EU", despite the providers being outside the EU). Yet it is unlikely to be enforced against providers from authoritarian regimes or anonymous users, which European authorities simply cannot reach.

Some hope lies in the introduction of provenance tools, using watermarks or hashing. Yet, ultimately, although these sorts of identifiers will help to protect authentic non-malicious content, they will do little to prevent intentional efforts to weaponize generative-AI content, as adversarial actors will simply evade such identification measures.

Democratic states can, however, take an important step towards preserving the integrity of information and shielding societies from the harmful impact of weaponized generative AI. In response to these challenges, the focus should shift towards developing effective mechanisms for detecting content produced by such models. This can be done by developing specific machine learning mechanisms trained to detect synthetically produced content. This is crucial to empower citizens to recognize generative AI content, enabling them to protect themselves from foreign – as well as domestic – propaganda and hateful content.

Given that the disruptive nature of generative AI is a concern for many governments – both democratic and authoritarian – the EU and the United States have put it on the agenda with other powers, and especially China, to explore whether there is some common denominator for co-operation, especially on authenticity issues. These can likely only be tackled if China, as a major producer of tech and AI, is ultimately part of the agreement.

Date: June 2023

This Rapid Response Brief was written by Jan Nicola Beyer, Digital Democracy Research Coordinator, and is part of Democracy Reporting International’s Disinfo Radar project, funded by the German Federal Foreign Office. Its contents do not necessarily represent the position of the German Federal Foreign Office.

About Democracy Reporting International

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