CSIS BRIEFS

Crossing the Deepfake Rubicon

The Maturing Synthetic Media Threat Landscape

By Di Cooke, Abby Edwards, Alexis Day, Devi Nair, Sophia Barkoff, and Katie Kelly

NOVEMBER 2024

THE ISSUE

- In recent years, threat actors have increasingly used synthetic media—digital content produced or manipulated by artificial intelligence (AI)—to enhance their deceptive activities, harming individuals and organizations worldwide with growing frequency.
- In addition, the weaponization of synthetic media has also begun to undermine people's trust in information integrity more widely, posing concerning implications for the stability and resilience of the U.S.'s information environment.
- At present, an individual's ability to recognize AI-generated content remains the primary defense against people falling prey to deceptively presented synthetic media.
- However, a recent experimental study by CSIS found that people are no longer able to reliably distinguish between authentic and AI-generated images, audio, and video sourced from publicly available tools.
- That human detection has ceased to be a reliable method for identifying synthetic media only heightens the dangers posed by the technology's misuse, underscoring the pressing need to implement alternative countermeasures to address this emerging threat.

INTRODUCTION

Synthetic media, which refers to text, images, audio, and video generated or manipulated by AI, presents both significant opportunities and risks. Recent advancements in generative AI technology have considerably **reduced** the data, computing power, and cost required to create highly realistic synthetic content. Coupled with the technology's growing accessibility, as evident from the rapidly **expanding** constellation of widely available user-friendly offerings, it has become easier than ever for anyone to manufacture genuine-seeming digital content using AI. The uses of such technology are seemingly endless, from the humorous, such as making fictional images of the pope **wearing** Balenciaga or videos of **Tom Cruise** dancing, to the commercial, such as streamlining work by assisting in email writing or creating digital avatars of people to use in training videos, news stories, or even for **speaking** with simulations of deceased loved ones. There has also been significant interest in harnessing generative AI's transformative potential for the greater good, from accelerating critical **scientific research** to making sophisticated **disability aids** like glasses that **translate** speech to text for the hard of hearing.

Yet, generative AI has also **become** a potent tool for misuse. On the morning of May 22, 2023, an AI-generated photograph reportedly **showing** an explosion near the Pentagon began to circulate extensively on social media platforms, causing widespread confusion and panic as well as a temporary but meaningful dip in the U.S. stock

Left: A synthetic image of Pope Francis wearing Balenciaga, which went viral on social media in 2023.[1]
Right: One of several synthetic videos featured on the now widely notorious Tom Cruise deepfake account on TikTok.[2]

market. While any adverse effects from this particular incident, in the end, were nominal, its occurrence nonetheless is illustrative of a broader trend of synthetic media being utilized to damaging ends. From criminal activities to adversarial military and intelligence operations, generative AI has more and more empowered the deception capabilities of threat actors, permitting them to manufacture convincingly realistic but fake digital content (colloquially known by many as "deepfakes") at unprecedented speed, scale, and degrees of precision. The rising ease of use and utility of the technology has led to a boom of AI-enabled deception incidents taking place over recent years, with the technology's abuse inflicting a growing amount of financial, reputational, physical, and mental harm to individuals and organizations worldwide. Already, the dangers posed by weaponized synthetic media have begun to shift from the theoretical to the realized.

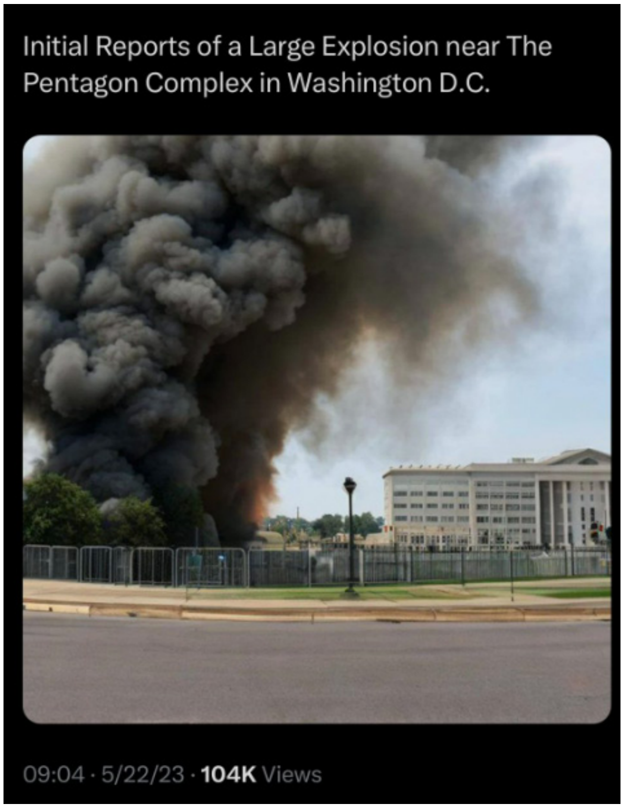
Thus far, the threat that has garnered the most public attention and alarm has been the **risk** of AI-enabled deceptions **disrupting** political elections by influencing voting outcomes, instigating unrest and violence, or damaging trust in the electoral process. Among the record number of elections held in 2024, the majority have already been subject to widely-circulated synthetic content that falsely **depicts** politicians or famous figures engaging in inappropriate or controversial behavior, criticizing their opposition, and promising policy changes. With the U.S. presidential election only a week away, widespread concerns have been understandably raised about the dangers of a flood of AI-generated content **amplifying** misinformation, or of an opportunely timed viral synthetic image or video of a political candidate influencing voters' decisions.

Today's synthetic media threat landscape extends far beyond the realm of political elections. AI-enabled financial fraud was found to have **risen** by 700 percent in 2023, and experts have predicted it will result in losses of up to $40 billion by 2027. Meanwhile, AI nonconsensual intimate media, accounting for 96 percent of all synthetic videos **online** as of 2019, has already claimed what is estimated to be **millions** of adults and children as victims – with that number **expected** to rise swiftly. Other AI-enabled deception incidents have also occurred with increasing frequency, spanning gray zone warfare such as influence operations and cyberattacks, espionage and surveillance, military deception operations, domestic disinformation, and more. As improvements in the technology's capabilities and accessibility continue, the volume and breadth of deception activity will likely grow.

The discrete harms arising from these incidents are further compounded by a more insidious danger: AI-enabled deception threatens to corrode the public's trust in the integrity of all information more broadly. There is already evidence that this has started to occur. In turn, this risks imperiling the foundations of the U.S.'s information environment, a vital pillar of societal stability and resilience.

Today, the principal defense against AI-enabled deceptions is people's ability to recognize synthetic media when encountering it in their day-to-day lives. However, rapid advancements in generative AI have increasingly constrained human detection capabilities as synthetic media has become more convincingly realistic. While the necessity of adopting alternative countermeasures, spanning from the technological to the regulatory, to compensate has been widely recognized as critical, in practice, implementation of these measures remains largely nascent. As such, this growing vulnerability means that awareness of when people are no longer able to depend solely on their eyes and ears to detect AI-generated content is critical in order to better recognize when human detection is no longer an effective safeguard against the technology's misuse.

To determine the current level of human detection capabilities, CSIS conducted a large-scale experimental **study** testing individuals' ability to differentiate between authentic media and synthetic images, audio, and videos sourced from publicly accessible generative AI technology.

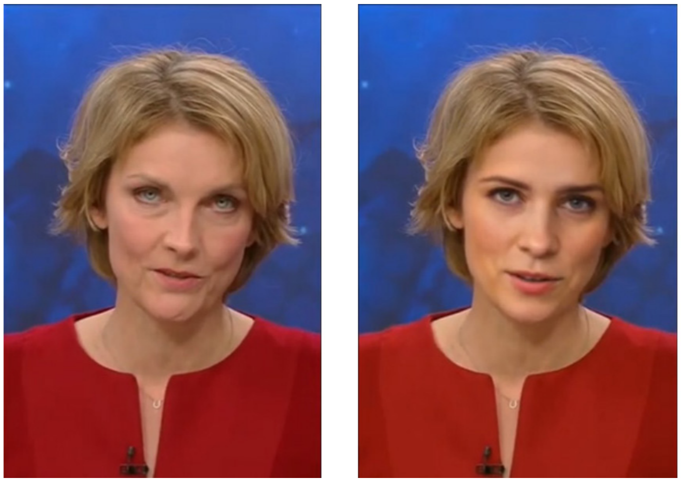


Initial Reports of a Large Explosion near The Pentagon Complex in Washington D.C.

09:04 · 5/22/23 · **104K** Views

This synthetic image was falsely reported as a photograph of an explosion near the Pentagon. It was widely circulated before being debunked as fake, causing widespread confusion and even a temporary dip in the U.S. stock market.[3]

Overall, the study found that people struggled to accurately identify AI-generated content to any meaningful degree, with some demographics being more susceptible to certain types of synthetic media than others. This brief reviews the study's key findings and offers an overview of the current synthetic media threat landscape, examining both ongoing and speculative harms in areas in which the abuse of this technology has become more prevalent. It is clear that weaponized synthetic media has begun to mature from an emergent to an established national security threat. That the inflection point has now been reached where human detection capabilities are unreliable only serves to underscore the pressing need to implement robust alternative countermeasures to address this growing danger.

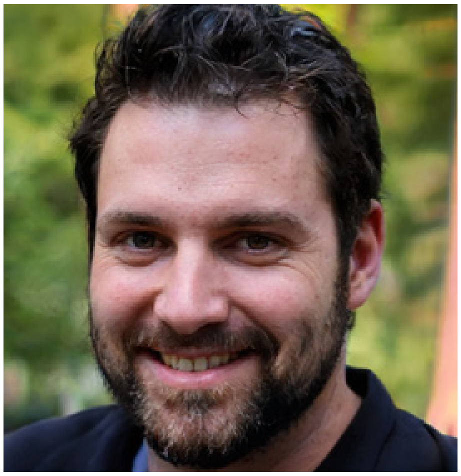


The study's most convincing synthetic audiovisual clip: When participants were presented with the AI-manipulated video clip (of comedian Nora Tschirner) on the right, 75.8 percent incorrectly labeled it as authentic. In comparison, the original video (of anchorwoman Marietta Slomka) is on the left. [4]

THE STUDY RESULTS

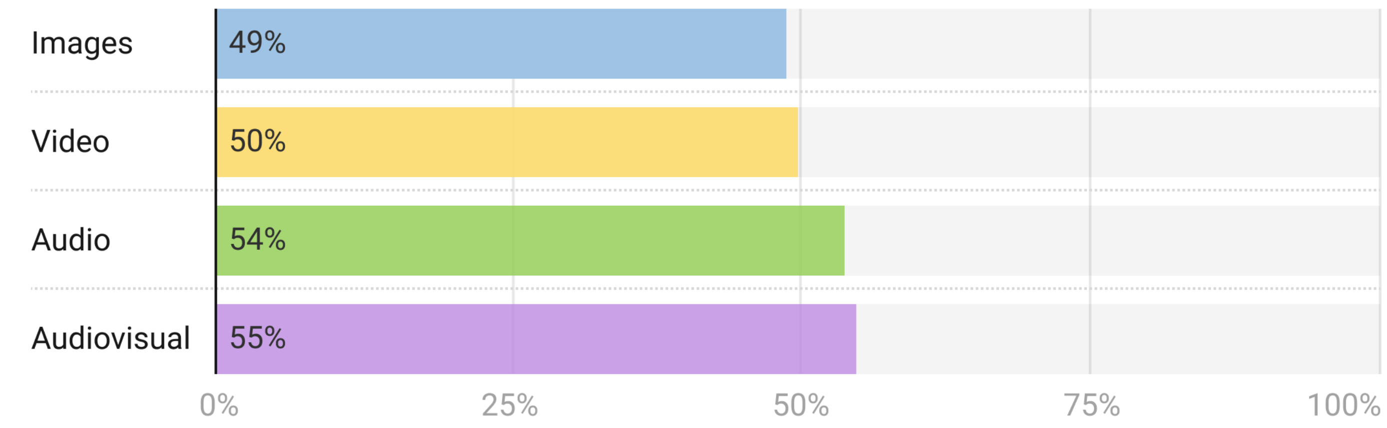
To assess how well people were able to detect AI-generated content, CSIS conducted a **perceptual study** involving nearly 1,300 North Americans aged from 18 to 85. Participants were asked to distinguish between synthetic and authentic media items, including images, audio, and videos both silent and fully audiovisual. The study also examined how other factors affected detection performance, including authenticity, language, modality, image subject matter, age, and participants' preexisting familiarity with synthetic media. To ensure that the AI-generated content would be representative of the quality and type of synthetic media people were likely to come across "in the wild", or in their daily lives, all synthetic test items were sourced from publicly available products and services.

Altogether, the study's findings paint a bleak picture of people's ability to discern the legitimacy of digital con-



The study's most convincing synthetic image: Only 10.7 percent of all participants correctly identified this as an AI-generated image, with the rest believing it to be a photograph of a real person.[5]

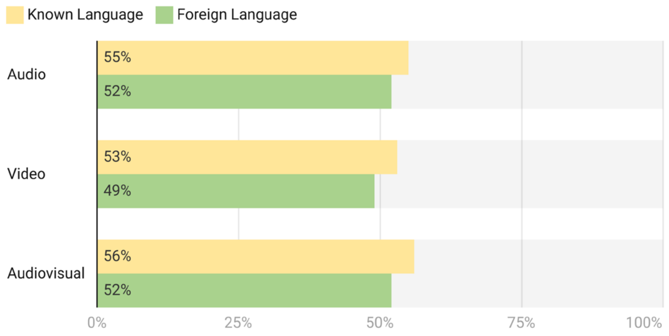
Figure 1: Average Detection Accuracy by Media Type



Source: Di Cooke et al., "As Good As A Coin Toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli," March 25, 2024, https://arxiv.org/abs/2403.16760.

tent in today's world. On average, participants correctly distinguished between synthetic and authentic media 51.2 percent of the time—roughly equivalent in accuracy to a coin toss. Images were the most difficult for participants to identify (49 percent average accuracy), with better detection performance on silent videos (51 percent) and audio clips (54 percent). Participants were the most successful at determining the authenticity of fully audiovisual clips (55 percent). These results are relatively unsurprising since public discourse and scientific research have closely **monitored** people's diminishing detection capabilities as generative AI has advanced in recent years. Nonetheless, it is valuable to confirm that this critical watershed moment has indeed been reached: humans can no longer depend solely on their own eyes and ears to reliably distinguish between reality and AI-generated falsehoods.

Figure 2: Detection Performance by Language Familiarity

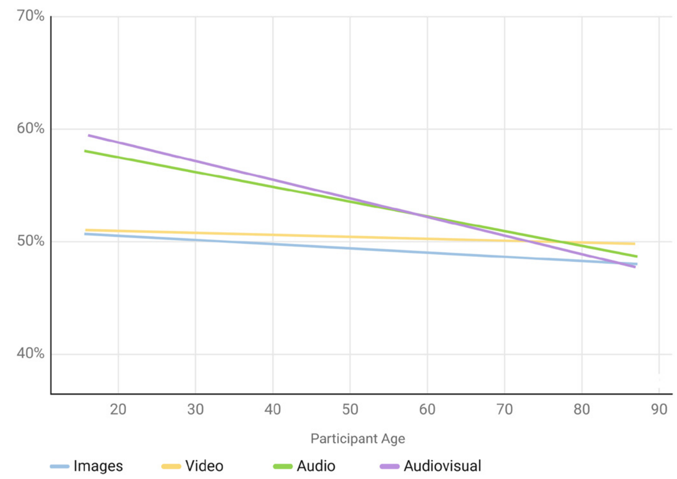


Source: Cooke et al., "As Good As A Coin Toss."

This does not mean that all AI-generated content being produced today is now indistinguishable from authentic media. Low- and mid-quality synthetic media still contains artifacts, or observable AI **glitches** such as bizarre-looking hands or illegible text, which make their provenance apparent. Regardless, our study demonstrates that numerous generative AI tools which are easily accessible to the public today can produce sufficiently realistic synthetic content that is relatively indistinguishable from authentic content to the human senses. Moreover, as the average quality of synthetic media improves while the technology matures, even low- and mid-quality outputs will become more realistic. For example, later iterations of AI image generators have already become increasingly capable of **rendering** real-looking hands, which makes relying on them as a potential "tell" of AI-generated content being present decreasingly useful.

The study's other findings offer more nuanced insights into people's vulnerabilities to different types of AI-enabled deceptions, examining how various elements may impact an individual's detection capabilities. For instance, participants' average detection accuracy was found to be significantly lower for audio, video, and audiovisual items featuring a foreign language than for items featuring languages in which they were fluent (see Figure 2). Meanwhile, younger participants outperformed their older counterparts to the greatest degree when tested on audiovisual and audio-only clips (see Figure 3). These findings indicate that people are more likely to misidentify synthetic media presented in a foreign language, and that older individuals are less sensitive to recognizing synthetic audio-based media.

Figure 3: Detection Performance by Age



Source: Cooke et al., "As Good As A Coin Toss."

Given the rise of multilingual synthetic misinformation as well as the growing popularity of AI phone scams which **often target** older generations, these findings suggest that these two demographics may be more vulnerable to certain types of AI-enabled deceptions than previously realized.

THE THREAT LANDSCAPE

The proliferation of weaponized synthetic media presents a clear and present danger to national security. To more effectively address these dangers, a more comprehensive understanding of the risks posed by its misuse and the various ways it has already been weaponized is required. Within only a few short years, the synthetic media threat landscape has expanded rapidly, with generative AI increasingly being exploited for nefarious purposes. Consequently, this rise in AI-enabled deception incidents has resulted in individuals and organizations around the world suffering financial, reputational, physical, and mental harm, even death, and countries worldwide experiencing detrimental effects on their societal stability and resilience. Now that it is clear publicly available generative AI tools can produce highly realistic synthetic media capable of deceiving even the most discerning of observers, these dangers have only become even more acute.

Generative AI has become an increasingly powerful force multiplier for deception, making it easier, faster, and cheaper to conduct more sophisticated stratagems than ever before – from producing synthetic content at an industrial **scale** to more precisely **tailoring** it to a target's specific vulnerabilities. These lowering barriers have, in turn, expanded the pool of threat actors who now are able to leverage this technology, from extremist **organizations** and organized crime **groups** to lone individuals with malicious intent. As of today, it costs less than $10 to create 30 minutes of customized synthetic **audio** featuring a target's voice or to manufacture a batch of over 1,000 individually personalized spear-phishing **emails**. Efforts to prevent the misuse of commercial products and services have been inconsistent in both their implementation and effectiveness, enabling the circumnavigation of guardrails to varying degrees of success. Meanwhile, open-source generative AI tools, which by their nature have more easily removable safeguards, have also furnished threat actors with a diverse and customizable toolkit, such as live **face-swapping** and **voice-masking** software, found to be used in real-time impersonation schemes. In addition, a shadow industry has begun to quickly develop to address this growing demand for purpose-built deception technologies. Spreading throughout the dark web and encrypted messaging **platforms**, it sells everything from prebuilt custom **software** to more bespoke **services** for explicitly abusive purposes.

However, the rise in AI-enabled deceptions has not been uniform. Rather, generative AI tools have been co-opted to greater degrees in scenarios where they currently provide a significant offensive edge to threat actors' stratagems over existing non-AI methods. For instance, the sharp rise of AI-enabled financial fraud over the past few years is a direct result of the substantial advantage afforded by AI technology, as AI text and audio generation tools are able to produce compelling synthetic content in less resource-intensive manners than when utilizing non-AI techniques. Conversely, AI-enabled deception incidents have been less prevalent in areas where synthetic media presently does not provide a similarly significant offensive edge. This has been **found** to be the case with deceptions involving the dissemination of false narratives, where conventional techniques such as manipulatively editing authentic media or sharing it out of context still remain highly effective and relatively easy to accomplish, limiting the comparative utility of generative AI tools. Regardless, as the technology's capabilities improve and barriers to using it decrease, it will undoubtedly be more extensively adopted for all manner of stratagems.

Compounding these harms is the second-order **risk** posed by weaponized synthetic media: the corrosion of information integrity. The proliferation of deceptive AI-generated content risks damaging the public's trust in the veracity of any information they encounter more generally as they become increasingly unable to trust their eyes and ears to reliably inform them as to what is real and what is fake. This degradation of trust in the truth **jeopardizes** the resilience of the U.S.'s information environment, or its '**epistemic security**' - which risks heightening its vulnerability to political and economic instability and constraining national security capabilities. Less epistemically secure societies are more **limited** in their ability to engage in collective and timely decision-making, making them more susceptible to adversarial manipulation, reducing their capacity for effective crisis response, and constraining critical defense and intelligence capabilities. This threat is not a novel one. Instances of widespread conventional misinformation have already been found to have diminished public trust in information from **media** and **government institutions**, resulting in **decreased** faith in political election integrity, **weakened** confidence in national security organizations, and **led** to violence and unrest. For example, pervasive false anti-vaccination narratives during the Covid-19 pandemic **undermined** vaccine confidence and institutional trust in the United States. The proliferation of these falsehoods, in turn, stymied economic growth, trade, and diplomacy, damaged education, and increased the number of vaccine-preventable outbreaks.

Synthetic media misuse risks intensifying the damage done to the public's trust in information by making it harder to distinguish fact from fiction. One can easily imagine how the viral AI-generated image of an **explosion** near the Pentagon, mentioned at the beginning of this brief, may have resulted in more significant adverse effects in a less epistemically secure society. Decreased public trust in information from institutional sources could have made later debunking by authorities less successful or take longer, enabling the falsehood to disseminate further and permitting greater knock-on effects to occur, such as more extensive financial volatility than just a brief dip in the stock market, which in turn could have led to civil unrest or facilitated the ability of foreign adversaries to leverage the unrest to their benefit.

There are signs that the increased prevalence of synthetic media has already begun to damage the public's epistemic trust. Research shows that repeated exposure to unlabelled synthetic media makes individuals **more** susceptible to misidentifying future synthetic content as well as **reduces** individuals' confidence in the truthfulness of all information. More recently, it was discovered that Russia's extensive use of AI-enabled deceptions throughout the still ongoing Russo-Ukrainian conflict has had a detrimental **effect** on Ukrainian citizens' confidence in information, making them significantly more skeptical of the truthfulness of all digital content they encounter online. Even just the existence of synthetic media itself has begun to erode aspects of the public's trust, as **evident** in the increasing frequency of authentic media being wrongly dismissed as AI-generated. The trend has become especially prevalent in information-contested spaces, such as political elections or the Israel-Hamas **conflict**, where both sides have frequently decried real digital content as being fake. As the synthetic media threat landscape continues to expand, these adverse effects will likely only grow stronger. Ultimately, it is the convergence of these immediate and systemic threats that makes countering weaponized synthetic media a national security imperative.

TYPES OF AI-ENABLED DECEPTIONS

The current synthetic media threat landscape can be broadly divided into six categories of AI-enabled deceptions: gray zone warfare, espionage and surveillance, military deception, domestic politics, nonconsensual intimate media, and financial crime. However, with the technology's continued advancement, it is anticipated that the depth and breadth of AI-enabled deception incidents will also expand and diversify, including **hate crimes**, falsification of evidence in **legal proceedings**, corporate **espionage** or sabotage, and more. To better illustrate the contours of today's landscape, a selection of particularly noteworthy AI deception incidents that have taken place across the six major categories have been shared below.

GRAY ZONE WARFARE

Synthetic media has been increasingly weaponized within **gray zone** warfare, or actions that take place in the murky waters between regular statecraft and outright warfare, such as information warfare, cyberattacks, and political and economic coercion. Examples of AI-enabled deception incidents that have occurred in the gray zone include the following:

- State-affiliated influence operations have disseminated synthetic images and videos as part of propaganda or information campaigns surrounding major political focal points or noteworthy events, including elections in **Europe** and **Taiwan**, **U.S. politics**, the **Russo-Ukrainian** and **Israeli-Hamas** conflicts, and the **2023 Maui wildfires**.

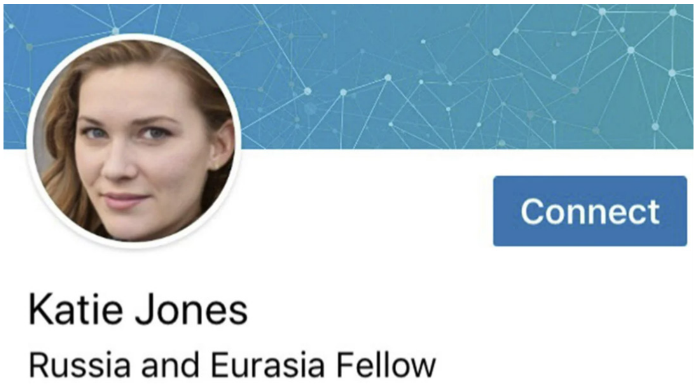
- Real-time face-swapping software was used to successfully **impersonate** Kyiv mayor Vitali Klitschko in a series of video calls with several mayors of major European cities as part of a targeted influence operation.
- An alleged AI-generated sex tape of a presidential candidate in the 2023 Turkish elections, **purportedly** published by an adversarial state, was widely circulated, leading to the candidate's withdrawal from the race.
- Chinese, Iranian, North Korean, and Russian state-affiliated actors were **found** to have been manufacturing synthetic content for spear phishing as part of cyberattacks they were planning to conduct.

Fake news channel clips featuring AI-generated TV anchors were shared by bot accounts online as a part of pro-Chinese information campaigns.[6]

ESPIONAGE AND SURVEILLANCE

Synthetic media has also been leveraged to a lesser extent, at least to public knowledge, for espionage and surveillance operations by states and the private cyber surveillance **industry**, strengthening online impersonations of real or fictitious individuals to obtain confidential information from targets. Examples of AI-enabled espionage and surveillance incidents include the following:

- Synthetic media was used in a fictitious Washington think tank employee's made-up **LinkedIn** account, which was suspected of being run as part of a Russian espionage operation.



Katie Jones

Russia and Eurasia Fellow

This fake LinkedIn account of a fictitious Washington think tank employee, suspected of being used for a Russian espionage operation, was found to have used a synthetic image for the profile photo.[7]

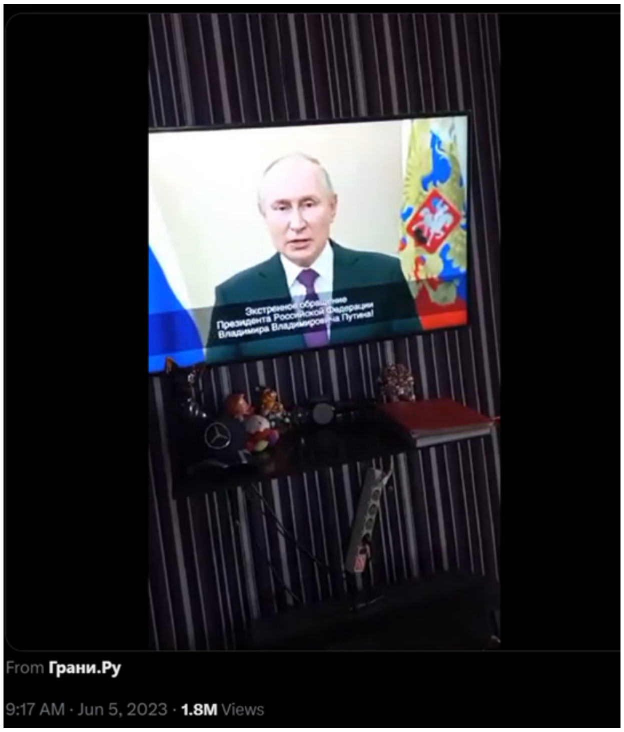
- Private cyber-intelligence companies used hundreds of fake accounts of social media content, **impersonating** activists, journalists, and young women, to covertly gather information from targets, including IP addresses and personal contact information.
- State-affiliated actors used social engineering assisted by large language models (LLM) to **manipulate** targets and facilitate the collection and analysis of open-source information.

MILITARY DECEPTION

Although the adversarial use of synthetic media for targeted military operations has so far been limited in practice, AI-enabled military deception remains a **topic** of great concern due to the large number of ways in which the technology could be leveraged to gain a battlefield **advantage**. This includes creating entirely fictitious events to alter or skew enemy intelligence, impersonating military personnel to falsify or muddle orders, and manufacturing **noise** to mask one's actions from an adversary or to overwhelm and confuse them. There are two particularly noteworthy examples of AI-enabled military deception incidents:

- AI-generated content featuring Ukrainian president Volodymyr Zelensky has been published and circulated extensively on social media to sow confusion and discord, including a synthetic **video** of him calling for his troops to immediately lay down their arms and surrender to Russian forces.

- Russian radio and TV networks were hacked to air fictitious AI-generated emergency **broadcasts** of Russian president Vladimir Putin **declaring** martial law due to Ukrainian forces invading Russian territory, causing some to actually evacuate in confusion.


Экстренное обращение
Президента Российской Федерации
Владимира Владимировича Путина!

From Грани.Ру

9:17 AM · Jun 5, 2023 · 1.8M Views

A social media post shows the airing of an emergency Russian TV broadcast featuring an AI-generated video that falsely depicted Russian president Vladimir Putin declaring martial law and calling for evacuation due to the Ukrainian invasion. [8]

DOMESTIC POLITICS

In recent years, there has been a **surge** in synthetic media being employed by domestic actors to create deceptive political content, **predominately** in regard to political elections. A selection of AI-enabled incidents include the following:

- The **Venezuelan** government **ran** fake news stories featuring AI-generated newscasters as part of a widespread domestic propaganda campaign to influence its citizens.
- Both pro-Israeli and pro-Palestinian social media accounts shared synthetic images of the ongoing **Gaza conflict**, such as AI-generated photos of a crying baby among bomb wreckage, to further false narratives.
- An AI-generated nonconsensual pornographic video of a **senior U.S. government official** at the Department of Homeland Security was circulated online as part of an ongoing smear campaign.
- Synthetic images of former president Donald Trump, portrayed as being real, were used in an **attack ad** by an opposition candidate during the U.S. presidential primaries.
- Synthetic media of politicians were falsely portrayed as authentic, including videos of UK prime minister **Keir Starmer** shouting at staff, U.S. president **Joe Biden** calling for a military draft, and a Slovakian presidential candidate discussing **vote rigging** during the election's final days.
- A **robocall campaign** used a synthetic audio clip of President Biden's voice to urge thousands of New Hampshire residents not to vote in the state's primary.
- U.K. far-right actors and politicians widely **circulated** anti-immigrant and Islamophobic synthetic content across social media ahead of the 2024 elections.
- Fictitious videos and images of celebrities such as **Taylor Swift**, as well as entirely AI-generated **Black voters**, endorsing former president Trump's 2024 U.S. presidential campaign have been frequently shared online by political supporters in the run up to the 2024 elections.

NONCONSENSUAL INTIMATE MEDIA

One of the most prolific abuses of generative AI to date has been the production of AI-generated **nonconsensual intimate media** of adults and children. Accounting for 96 percent of all synthetic media **videos** in existence in 2019, the adult nonconsensual pornography **industry** and the **online** trafficking of child sexual abuse materials have **exploded** in the years since, claiming millions of adult and child victims to date. Examples of incidents include the following:

- An Indian journalist investigating the rape of a young girl was **the target** of an extensive hate campaign, which included synthetic pornography of her being circulated online.
- An automated **Telegram** bot service created and published sexual images of an estimated 24,000 women and girls.
- Sexually explicit images and videos of school girls and female teachers being produced and shared

online by male students in **Korea**, **Brazil**, **Spain**, and the **United States**, among others.

- A **recently** uncovered worldwide trafficking ring producing and selling sexually synthetic images depicting photorealistic children on a reported "industrial scale."
- A **deluge** of synthetic pornography featuring Taylor Swift spread across the social media platform X, forcing the online platform to block searches of the celebrity temporarily.

FINANCIAL CRIME

AI-enabled **financial crime** has quickly become one of the most widespread misuses of synthetic media. Criminals have employed generative AI tools to **impersonate**, **extort**, and **hack** for a multitude of fraudulent activities, with personalized AI spear-phishing emails and voice phone scams experiencing the largest growth. With an estimated 700 percent **increase** in incidents in 2023 from the previous year, financial experts predict that AI-enabled financial fraud could lead to **losses** of $40 billion by 2027. Noteworthy incidents include the following:

- The head of a UK energy firm was personally **tricked** into transferring nearly $250,000 by fraudsters who used voice cloning to impersonate the parent company's CEO.
- Reportedly the largest AI-generated scam to date, thousands of synthetic videos of **celebrities** such as **Elon Musk** and MrBeast promoting fake financial schemes have been widely circulated on social media platforms.
- An Arizona woman was the **target** of a fake ransoming scheme in which fraudsters impersonated her daughter over the phone using voice-cloning technology.
- An employee of a financial firm was tricked during a week-long ruse into paying out $25 million to fraudsters after the scammers used real-time synthetic audiovisual software to **impersonate** the employee's senior personnel and colleagues through a series of group video conferences, emails, and calls.
- The Yahoo Boys, a **crime collective**, have widely adopted AI tools for romance scams and sextortion, employing live face and voice impersonation software and "nudification" apps to trick and blackmail targets. This has led not only to financial loss but also to tragic deaths in which some targets, frequently teenagers, took their own lives.

CONCLUSION

As generative AI technology continues to advance, so does the potential for its misuse. In only a few short years, the synthetic media threat landscape has changed dramatically. AI-enabled deceptions have become increasingly complex and varied, ranging from gray zone warfare to financial fraud and beyond. Not only has the weaponization of synthetic media already begun to cause real and substantial harm to people and organizations worldwide, but it also threatens to undermine public trust in all information online, regardless of the truth. Overall, these developments present troubling implications for U.S. national security.

These dangers have become even more severe as it has been made clear that widely available generative AI technology has progressed to the point that people can no longer depend on their eyes and ears to reliably detect the synthetic content they might encounter in their everyday lives. With this primary line of defense compromised, pursuing alternative solutions has never been so vital. Now more than ever, stakeholders across the private and public sectors must work together to implement multifaceted countermeasures that bridge the technological, regulatory, and educational domains to oppose the growing threat posed by weaponized synthetic media. ■

Di Cooke *is a horizon fellow with the International Security Program at the Center for Strategic and International Studies (CSIS) in Washington, D.C.* ***Abby Edwards*** *is a former research associate in the International Security Program at CSIS.* ***Alexis Day*** *is an associate director for the Smart Women, Smart Power Initiative at CSIS.* ***Devi Nair*** *is a former associate director and associate fellow in the International Security Program at CSIS.* ***Sophia Barkoff*** *is a former research intern in Defending Democratic Institutions in the International Security Program at CSIS.* ***Katie Kelly*** *is a former social media and outreach intern in the International Security Program at CSIS.*

The authors would like to thank Matthew O'Shaughnessy, Sid Srinivasan, Daniel Byman, Caitlin Chin-Rothman, Hadrien

Pouget, Elizabeth Sagar, Guilio Corsi, Max Katz, and countless others for their valuable feedback and insights.
They would also like to thank Catrina Doxsee, Riley McCabe, and Gamin Min from the Transnational Threats Project at CSIS for their help in beta testing and Gamin Min, Ike Barrash, and Daniel Pycock for their data analysis support. Finally, they would like to thank everyone who participated in the pilot survey.

This report was made possible by general support to the CSIS International Security Program. No direct sponsorship contributed to this report.

Cover Photo: Di Studio | gsshot | andriano_cz/Adobe Stock

ENDNOTES

[1] Pablo Xavier, “The Pope Drip,” Reddit, March 24, 2023, https://www.reddit.com/r/midjourney/comments/120vhdc/the_pope_drip/.

[2] deeptomcruise, TikTok video, December 26, 2022, 00:19, https://www.tiktok.com/@deeptomcruise/video/7181490100314885382?lang=en.

[3] OSINTdefender (@sentdefender), X post, May 22, 2023, 09:04 am, https://x.com/sentdefender/status/1660650575569059840/photo/1.

[4] Deepfacelabfan, “Deepfake - Marietta Slomka zu Nora Tschirner - 128 LIAE 15k RW only - 90min FAKE,” YouTube video, April 7, 2022, 00:46, https://www.youtube.com/watch?v=V4ln4SyVN-jg&list=PL8ax9s9DVKClTiPm9c5Wkq9wOG4hGKyVH&index=5.

[5] Di Cooke et al., “As Good As A Coin Toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli,” March 25, 2024, https://arxiv.org/abs/2403.16760.

[6] Adam Satariano and Paul Mozur, “The People Onscreen Are Fake. The Disinformation Is Real.,” *New York Times*, February 7, 2022, https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html.

[7] Raphael Satter, “Experts: Spy used AI-generated face to connect with targets,” AP News, June 13, 2019, https://apnews.com/article/ap-top-news-artificial-intelligence-social-platforms-think-tanks-politics-bc2f19097a4c4fffaa00de6770b8a60d.

[8] Alex Kokcharov (@Alex Kokcharov), X post, June 15, 2023, 6:17 am, https://x.com/AlexKokcharov/status/1665709387648827397.