Developing a Disinformation Incident Response Playbook:

Combatting Real-Time Disruption via Deepfakes and Generative AI

Edward L. Mienie, University of North Georgia

Bryson R. Payne, University of North Georgia

Abstract

We are entering an age in which disinformation, fake news, and falsified images, videos, and audio 1) are rapidly becoming indistinguishable from authentic media, 2) can be produced in real-time, and 3) can be deployed at scale and in quantities that businesses, news agencies, and nations alike may not be able to respond to effectively. Over the past four years, deepfake videos, in which an actor's face can be replaced with a believable facsimile of a CEO's or other famous person's face, have become relatively commonplace in popular culture, and deepfakes have already been used at least at a rudimentary level in disinformation campaigns. ChatGPT, a generative large-language AI model that can produce authentic-sounding human-readable text, can generate fake news articles, emails, and blog or social media posts in real-time that seem fluent and realistic to the reader. Newer generative AI tools for creating audio, video, and photorealistic images can lend additional credibility to disinformation, misinformation, and fake news and spread them online faster than human reporters and government officials can fact-check or respond. This research examines the perfect storm of disinformation enabled by these combined technologies, provides a review of existing and emerging literature in the field, and includes a brief case study on Ukraine's response to the 2022 Zelensky deepfake video at the onset of the Russian invasion to draw out recommendations for businesses, governments, and news organizations in countering AI-enhanced disruption.

Keywords: cybersecurity, identity theft, account security, multi-factor authentication

Introduction

"Falsehoods traverse the globe while veracity is still fastening its trousers."
– ChatGPT, 2023

Artificial Intelligence (AI) is a technological innovation with the potential to disrupt most aspects of human life. The era of fake news, misinformation, disinformation, and post-truth is already impacting decision-making within the realm of organizational and national security. With generative AI capable of producing realistic text, speech, images, and audio through such technologies as ChatGPT, Stable Diffusion, Midjourney, and others, added to existing deepfakes video-altering technology, the line between reality and machine-generated misinformation has not only blurred, it has been all but erased.

A recent study by University College London found that both English and Mandarin Chinese speakers were able to correctly identify artificially generated speech only 73% of the time at 2023 levels of technology (Mai, Brai, Davies & Griffin, 2023). This means that with now-dated AI speech generation technology, 27% of users would likely believe the content of AI-generated audio was authentic. Furthermore, researchers have demonstrated that humans mistake deepfake and authentic videos as much as 66% of the time even when the two are shown side-by-side (Allen et al., 2023), and technology in this area is expected to continue to advance, making it even more difficult to discern true human speech and video from completely fabricated AI-generated media (Farah, 2023).

Newer generative AI tools are capable of creating believable audio, video, and photorealistic images that can be used to spread misinformation and disinformation faster than public relations and government officials can respond. This research outlines the converging perfect storm of disinformation enabled by these combined technologies, provides a review of existing and emerging literature related to AI-generated and AI-enhanced disinformation, and provides a brief case study on Ukraine's response to the 2022 Zelensky deepfake video at the onset of the Russian invasion. The goal of this research is to compile recommendations for businesses, governments, and even individuals in developing a disinformation incident response playbook to counter AI-enhanced disruption in real-time.

Organizational and National Security Implications

Rapidly advancing generative AI technology can reasonably have a disruptive effect on the decision-making process, both for individuals and for businesses, as well as for governments and national security. Decision makers already must contend with changing ideologies, policies, other disruptive technologies, cultural shifts, and social changes in addition to traditional adversarial threats. Artificial disruption in near real-time adds another challenging dimension to an existing and increasingly complicated, duplicitous, and counterfeit world. Deepfakes, ChatGPT, and generative artificial intelligence (a subset of machine learning, or ML) pose challenges to the established mechanisms used to inform decision-makers about world events. Ultimately, decision-makers must make the most informed and accurate decisions that will enhance our nation's national security. Today, national security advisors are being challenged by the ever-changing and increasingly sophisticated daily advancements of AI. In the major fields of AI, quantum technologies, and advanced materials, China is the leading country in 37 of 44 technologies, producing more than five times as much high-impact research as the US, its closest competitor (ASPI, 2023).

The federal government has identified AI's possible applications for defense and intelligence and has made it a major priority. However, Tucker (2020) argues that policymakers and leaders must better understand how AI systems reach their conclusions, and before the United States Intelligence Community (IC) can use AI to its full potential, it must be hardened against attack. The Office of the Director of National Intelligence (ODNI) in 2018 launched a strategy for augmenting intelligence using machines (AIM) to foster stronger collaboration by organizing and sharing their AI efforts thereby creating a synergy across the IC. This initiative forces agencies to look outside their environments as they are so consumed within their spaces with these fast-developing AI systems. ODNI is keen to integrate the IC's many unintended information silos, and agencies are applying a more integrated approach to AI to help transform tradecraft (Shapiro, 2022). "We're really working toward a whole-agency approach toward AI," Lakshmi Raman, the CIA's chief of AI said at a recent Intelligence and National Security Alliance conference. She stated that AI technology has "relevance for data collection, analysis, digital innovation, operations, and even legal and finance areas" (Shapiro, 2022). Some of our main adversaries are also investing in AI research and development, and in our collective quest for competitive advantage, the potential to use poorly understood or untested systems could lead to serious unintended consequences that could impact the world.

The business world is wise to embrace AI technologies, as research shows that business organizations become more competitive, efficient, and innovative when doing so. However, if AI is going to serve the good of humanity, it must be applied responsibly and ethically if it is going to be the key driver for positive change as anticipated (Martinovic, Bandur, & Tusevski, 2024).

Implications for the Intelligence Community

The United States' intelligence community (IC) has been focused on AI for a long time, examining ways to leverage its power, and, by implication, give the US an advantage to set precedents that other international actors could resist, comply with, or negotiate (Moran, Burton, Christou, 2023). The Defense Advanced Research Projects Agency (DARPA) announced a $2 billion campaign to develop the next wave of AI Technologies to research more collaborative and trusting partnerships between machines and humans (DARPA, 2018). The IC recognizes that the private sector performs the lion's share of AI systems research and development and that working with the private sector poses challenges of many sorts, including conflicting interests and ideologies (Moran, et al, 2023). Moran et al (2023) posit that it is premature to talk about an intelligence revolution brought about by AI because of cultural tensions within the global AI ecosystem and local and international rules and regulations governing data collection and storage.

In 2022, The White House announced that it wanted a "Blueprint for an AI Bill of Rights," which should "protect the American people from unsafe and ineffective systems;" that they should "not face discrimination by algorithms and systems should be used and designed in an equitable way;" that they should be "protected from abusive data practices via built-in protections and should have agency over how data about them is being used;" that they "should know that an automated system is being used and understand how and why it contributes outcomes that impact them;" and that they "should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems they encounter" (The White House, 2022). The White House in May 2023 pledged a "road map" for managing AI. The plan provides for "international cooperation to manage the impact of AI." The White House acknowledges the broad applications of AI, while at the same time recognizing that the risks that AI presents need to be managed effectively by way of regulatory intervention by governments worldwide (Milligan, 2023).

At the same time, we realize that AI technology is growing at a rate faster than regulators can respond to it. The Bipartisan Policy Center and Georgetown University's Center for Security and Emergency Technology posit that, inter alia, the US must work closely with its allies and partners while also cooperating pragmatically and selectively with its adversaries such as Russia and China; prevent the transfer of sensitive AI technologies to China through export and investment controls; and implement processes to develop and deploy defense and intelligence applications of AI systems by focusing on trustworthiness, human-machine teaming, and Department of Defense's (DOD) ethical principles for AI (Bipartisan Policy Center, 2020).

AI presents an array of opportunities for strengthening the efficiency and effectiveness of intelligence procedures and challenges to the organizational structure within the U.S. Intelligence Community (IC). AI capabilities could be 1) integrated into the collection, analysis, and management processes, 2) used to educate the IC workforce as AI will influence world political events, and 3) assist the IC in the collection, analysis, and dissemination of information on AI developments worldwide. The IC would have to dedicate time, resources, and effort to accomplish the aforementioned (Blais & Jungdahl, 2019). AI can be used to augment the IC's intelligence efforts and not replace it, thereby supporting and not determining the decision-making process to enable them to make better decisions. The human element ultimately should remain supreme even if it's just for cognitive and critical thinking purposes that can help to process nuances as no machine can...yet. Bias and discrimination in developing AI systems pose challenges to producing accurate results, and countering this would require highly developed processes and specialist expertise (U.K. Government, n.d.). We must maintain an active and engaged human dimension that is capable of processing nuances of the ever-changing, asymmetrical, network-centric world (Faunt & Gentile, 2019).

Some of the positive business implications for using AI responsibly are improved decision-making and customer service, stimulation of new ideas, automation, creation of new opportunities in the marketplace, and gaining valuable insights by analyzing massive volumes of data thereby enabling these organizations to make better-informed forecasts. AI can process complex data thereby giving organizations better insights to enhance efficient business practices (Martinovic, Bandur, & Tusevski, 2024). As AI can collect vast amounts of personal data which may raise ethical and privacy issues, businesses should mitigate the threat that cybercrime poses to protect such sensitive personal information.

CyberCrime and Cyber Warfare Using AI

Deception Using Deepfakes

The threat posed by real-time AI-manipulated media endangers both businesses and nation-states. Imagine taking a Zoom call from your CEO, asking you to transfer money to address an urgent business matter, but later finding out it was a cybercriminal using both video and voice deepfake technology in real-time to steal from your organization. Zror (2023) used this technology live, on-stage at the international hacker conference DEFCON, to impersonate the conference's founder Jeff Moss and state that "DEFCON is canceled" in front of nearly 30,000 attendees.

Ukrainians able to access the web in March 2022 saw a video clip of President Volodymyr Zelensky, speaking behind a podium with the Ukrainian state seal behind him in his usual battle fatigues, making a call to all the Ukrainian soldiers to lay down their arms and to return to their homes (The Guardian, 2022), as shown in Figure 1.

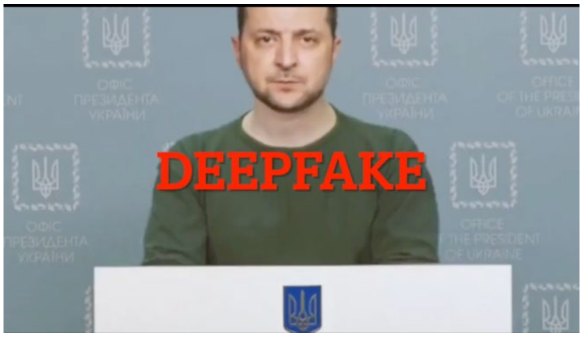


Figure 1: A deepfake video of Ukrainian President Volodymyr Zelensky appearing to urge Ukrainians to surrender to the Russian invasion in early 2022. Mikael Thalen (@MikaelThalen) Twitter, March 16, 2022, https://twitter.com/MikaelThalen/status/1504123674516885507.

Such a sophisticated hoax can have serious national security implications if taken as authentic. While sources are unsure whether the deepfake video of Zelensky released shortly after the Russian invasion of Ukraine originated in Russia, the message's intent seemed clear in asking Ukrainians to surrender to the Russian invasion. Fortunately for Ukraine, Zelensky and his team responded to and successfully debunked the deepfake within 24 hours of its appearance in world media (Simonite, 2022).

Manipulating multimedia has been made possible through the progress in machine learning, AI, and deep learning that has produced new tools and techniques. Today we have powerful tools such as the generative adversarial network (GAN) framework that assists with the production of high-resolution photorealistic videos and images. In 2019 Israel arrested three Franco-Israeli conmen who impersonated the French Foreign Minister after defrauding a businessman out of eight million Euros (The Guardian, 2019). GAN could be applied to image processing, image translation, and video synthesis (Liu, Huang, Yu, Wang & Mallya, 2021), and it should be noted that people have been blackmailed, harassed, and political discord and hate have been incited through realistic fake and high-quality images, videos, and audios.

Deepfakes are manifested as high-quality, manipulated videos, a product of machine-learning applications that create a fake video that otherwise appears authentic by combining, replacing, merging, and superimposing video clips and images onto a video (Maras & Alexandrou, 2019). Rana, Nobi, Murali & Sung (2022) identified various approaches to tackle the challenges brought on through deepfakes and categorized them as: (1) deep learning-based techniques, (2) classical machine learning-based methods, (3) statistical techniques, and (4) blockchain-based techniques. They conclude that deep learning-based methods are by far the best in tackling the challenges deepfakes pose to decision-makers. In the end, as technology improves, efforts will need to be made to identify and expose deepfakes by developing parallel technologies to detect them.

ChatGPT as an Instrument of Misinformation

The large language model (LLM) known as ChatGPT has been making headlines since November 2022 due to the release of an advanced version capable of producing highly realistic AI-generated text that sounds like it could have been written by a human. Many cases nationwide cited the use of ChatGPT in producing homework essays that could fool college professors, and it has even been able to write legal briefs capable of fooling at least one lawyer, who is now facing sanctions for using the AI system to generate fake case law (Maruf, 2023).

In two additional cases we see that it is not a question of what information AI reveals, but more a question of what it is algorithmically programmed not to reveal. One such example ties in with Russia’s control over the media, prohibiting any negative information being published about their leader, Putin. Another example is what the Chinese Communist Party allows LLMs to publish around the persecution of Uyghurs and the Tiananmen massacre (Urman & Makhortykh, 2023). We also need to be aware of the potential discrimination and biases that may be reflected in the vast data that could contribute to the ethical challenges that are associated with ChatGPT. Propagating and generating misleading information is an ethical concern (Kareem, 2024).

ChatGPT has been used to develop social engineering phishing emails designed to trick people into revealing their banking or other sensitive information, and related technologies have been used to harass individuals and endanger lives by “swatting” or calling police SWAT tactical teams with false information about non-existent threats at real addresses and homes (Cox, 2023).

As an example of using ChatGPT to generate fake news stories for major media outlets, the authors envisioned a scenario in which China might reassert its physical control over Taiwan and asked ChatGPT to write up the event in a Reuters-styled press release. The result was a believable, AI-fabricated news story that began as follows:

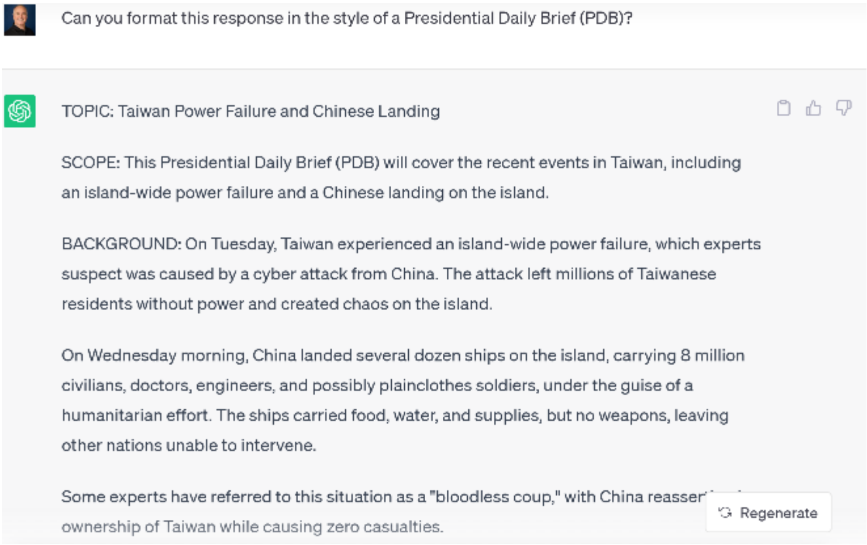
> *Taipei, Taiwan - Taiwan experienced an island-wide power failure on Tuesday, with experts suspecting that China used a cyber attack to black out the island. The attack left millions of Taiwanese residents in the dark and created chaos on the island. On Wednesday morning, China landed several dozen ships with 8 million civilians, doctors, engineers, and possibly plainclothes soldiers in what it labeled as a humanitarian effort to help its*

> *"little brother". The arriving ships carried food, water, and supplies, with no weapons, leaving other nations unable to intervene. While the move was initially seen as a helpful gesture, some experts have referred to it as a "bloodless coup," with China reasserting its ownership of Taiwan while causing zero casualties...*

The press release ended with the following statements, again completely AI-generated:

> *…Many fear that the situation could escalate further, raising tensions between China and the United States, and potentially leading to a larger conflict in the region. As of now, it remains to be seen what the future holds for Taiwan and its relationship with China. The situation is still unfolding, and many are watching closely to see how events will unfold in the coming days and weeks.*

Perhaps a larger concern is the speed with which such stories can be generated, in a variety of formats. In less than 5 seconds, the same story was reformatted as a Presidential Daily Brief (or PDB), a format commonly used in the US Intelligence Community. A screenshot of ChatGPT's AI-generated presidential briefing is shown in Figure 2.



Can you format this response in the style of a Presidential Daily Brief (PDB)?

TOPIC: Taiwan Power Failure and Chinese Landing

SCOPE: This Presidential Daily Brief (PDB) will cover the recent events in Taiwan, including an island-wide power failure and a Chinese landing on the island.

BACKGROUND: On Tuesday, Taiwan experienced an island-wide power failure, which experts suspect was caused by a cyber attack from China. The attack left millions of Taiwanese residents without power and created chaos on the island.

On Wednesday morning, China landed several dozen ships on the island, carrying 8 million civilians, doctors, engineers, and possibly plainclothes soldiers, under the guise of a humanitarian effort. The ships carried food, water, and supplies, but no weapons, leaving other nations unable to intervene.

Some experts have referred to this situation as a "bloodless coup," with China reasser
ownership of Taiwan while causing zero casualties.

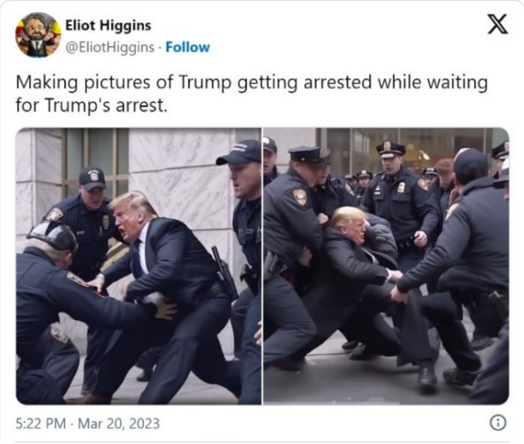
Figure 2: A fake press release from ChatGPT describing the takeover of Taiwan by China after an apparent cyberattack leading to an island-wide electrical grid blackout, formatted as a Presidential Daily Brief (PDB), a high-level US intelligence format used in the White House.

The PDB version used slightly different wording, with more detail but less prose, along with recommendations for action at the end. Using similar prompts, a small team of information warfare or psychological operations special operatives within any nation could generate literally hundreds of unique but corroborating accounts of a fictitious scenario of a similar or greater magnitude and post them to news sites or individual social media accounts, lending credibility and creating confusion regionally or world-wide.

The Generative AI Threat

An emerging threat to cybersecurity is the use of generative AI by hackers. Attackers generate fake videos, audio, images, and text by using generative AI, which allows them to launch cyberattacks such as social engineering, phishing scams, password cracking, impersonation attacks, malware development, and others. In 2019, criminals impersonated a chief executive's voice and demanded a fraudulent $243,000 transfer using AI-based voice-spoofing attack software (The Wall Street Journal, 2019), and a mayoral candidate in Chicago was impersonated by AI audio online making inflammatory remarks (Kahn, 2023), demonstrating the potential for future election misinformation to be spread using AI-generated content.

Another recent example of generative AI was used against 2024 US presidential candidate Donald Trump (Lu, 2023). The Midjourney AI image generator was used by journalist Eliot Higgins to depict Mr. Trump being arrested, at a time when Trump was being indicted on federal charges. Higgins indicated when posting the images on Twitter that he had generated the pictures using AI, but some of the pictures were captured by foreign media and presented without the information that the images were false and AI-generated (Di Placido, 2023). The image of a leading candidate from a major political party being tackled and dragged away by police (Figure 3a) is an extreme example of the kind of misinformation that generative AI can produce given even a simple query string. A more positive but still improbable image was generated by the authors by typing the query phrase "Donald Trump and Joe Biden holding hands in victory on a campaign stage with American flags," into the generative AI website, Stablediffusionweb.com (Figure 3b).

a) 

Eliot Higgins
@EliotHiggins · Follow

Making pictures of Trump getting arrested while waiting for Trump's arrest.

5:22 PM · Mar 20, 2023

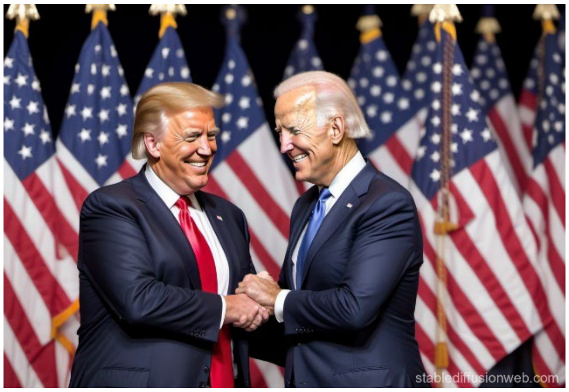
b) 

Figure 3a. Midjourney generative AI produced realistic-looking pictures of US presidential candidate Donald Trump being arrested, at a time when the candidate was being indicted for federal crimes. (Source: https://twitter.com/EliotHiggins/status/1637927681734987777).
Figure 3b: Stable Diffusion AI generated this slightly more light-hearted disinformation with the query "Donald Trump and Joe Biden holding hands in victory on a campaign stage with American flags." (Source: stablediffusionweb.com)

Hackers have the option of using different types of generative AI, which generates new data or multimedia content that is difficult to distinguish from authentic human-sourced text, images, audio, or video. Present generative AI systems are typically based on generative adversarial networks (GANs), variational autoencoders (VAEs), and recurrent neural networks (RNNs) (Biniyaz, 2023). Microsoft co-founder, Bill Gates, acknowledges that AI has the potential to impact society in very profound ways, while Elon Musk, Tesla CEO, has said that AI is "actually far more dangerous than nukes" (Clifford, 2019). According to technology and AI experts, the dangers of AI, if used nefariously in digital, political, and physical attacks, could include analysis of human behaviors, speech synthesis for impersonation, moods and beliefs for manipulation, and physical weapons such as micro-drones and other physical weapons (Clifford, 2018, 2019).

Additional concerns include threats to intellectual property/copyright, security, privacy, and considerations of discrimination and bias in LLMs' output. Scams and misinformation efforts may arise because of the illegal exploitation of gen AI. AI prizes plausibility over accuracy as was shown in a 2023 court filing by a Georgia radio host against ChatGPT who falsely stated that he embezzled funds from another organization (Isik, Joshi, & Goutas, 2024). Misrepresentation by a third party of another's product can question the credibility of that product as was the case of a Tesla cybertruck crash as portrayed in a deepfake video (Isik, Joshi, & Goutas, 2024). Unaware of inauthentic content, users may inadvertently share that content, which could negatively impact the shareholder value of an organization (Isik, Joshi, & Goutas, 2024).

Application of A.I. in Information Operations

Covert action is used to achieve foreign policy objectives by influencing the way the target audience thinks and believes. Such action is sometimes referred to as information operations, propaganda, or psychological operations. By way of an example, the Soviet Union launched a fake news campaign near the end of the 20th century by spreading false rumors that the Acquired Immune Deficiency Syndrome (AIDS) disease was a US biological weapon (White House, 2022). We know that the USSR used "dezinformatsiya" (disinformation) during the Cold War to further its foreign policy objectives by sowing division and fear among its target audience. In this case, they used India's news media to plant the fake story. Had the USSR had access to AI at that time in the 1980s, this fake story would have instantly permeated social media platforms in addition to the traditional media outlets worldwide, and the intended effect of discrediting the US could have been perceived to be significantly more plausible. This fake story, at least initially, may have enjoyed credibility for a longer time causing the US to increase its efforts of discrediting the fake story, thereby distracting the US from more pressing issues of national security.

In another example of covert action at work on the political front, shortly after World War II, the Italian Communist Party threatened to oust the Christian Democratic government. To counter the surge of money coming from foreign sponsors in support of the Communists, the US had to raise funds among the anti-communist labor unions and the Italian American community to counter those efforts. Through a letter-writing campaign launched by the US government via Italian Americans to their fellow Italians, an effort was made to convince the Italians that life under a capitalist system is much better than under a communist system (Acuff et al., 2022). Had AI existed at that time, this information operation could have reached a much wider audience in nanoseconds via various social media platforms and resulted in a significant impact on the voters not to support the Communist Party.

Another area of covert action is the removal of a dangerous foreign leader through the fomenting of a *coup d'état* in place of waging war against that country. An example of such covert support to a *coup* was the removal of the socialist president of Chile, Salvador Allende, in the early 1970s. Unfortunately, he was replaced by Augusto Pinochet, who ran a brutal military regime until 1990. The CIA denied any involvement in the coup, although it is widely believed that the US Government was behind the plot (Acuff et al., 2022). Had AI existed at that time, it would have been easier for the US government to proffer plausible deniability by waging an information campaign via social media and other media platforms aimed at discrediting those rumors and allegations in real-time.

Sabotage does not only manifest as violence, but it could also be non-violent in nature if it is intended to foment instability through information operations. An example of such non-violent means was Russian covert acts to sow dissension and confusion among Americans when it was aimed at Hillary Clinton's presidential bid in 2016 (Acuff et al., 2022). If AI had been developed where deepfakes could have been used in the Russian covert operations, it would have made their operation more plausible and effective by impersonating her and other opposition candidates' fake responses to enhance even more division among the populace.

Today, Russia appears to be embracing a new model of sabotage referred to as the so-called gig economy, one which focuses on a temporary workforce made up of freelance contractors in a largely free-market online system to support Putin's plausible deniability efforts when nefarious cyber operations are conducted against enemy states (Richterova et al., 2024). It appears that Russia's current sabotage operations mostly resemble the crowdsourced type, used for recruitment on Telegram by Russian intelligence services (Richterova et al., 2024).

Covertly facilitating strikes that target a large company, an industry, or a country's economy would be another example of sabotage complemented by a disinformation campaign that could undermine investor and consumer confidence. Having access to AI platforms can enhance the messaging to reach a wider intended audience in a much shorter time thereby amplifying the intended effect of undermining the economy. Fake imagery, audio, and video of major economic and political players, along with a tidal wave of fake news stories, social media posts, and even disinformation spread by AI influencers—fake or licensed personas where all content is generated by AI—can, and likely will, be combined into a perfect storm of disinformation, distraction, and disruption.

Furthermore, cyber financial sabotage against a nation or large corporation can potentially deter investment, disrupt economic activities, and undermine confidence in the stability of the company or financial system. Currency manipulation and economic sanctions may lead to capital flight, market volatility, and recessionary pressures with their concomitant negative effects on growth and economic instability, which may exacerbate poverty and social inequality (Green, 2024).

Responding to Real-Time Disinformation: A Brief Case Study

How, then, can businesses, organizations, and governments plan for and respond to the emerging threat of real-time information warfare from AI-enhanced adversaries including nation-states, terrorist groups, organized crime, and rogue individuals? The Zelensky case study offers a tested approach: developing a playbook for disinformation incident response. This playbook is similar to cybersecurity incident response plans developed over the past three to four decades for businesses and governments, plans now considered crucial for compliance and business continuity/disaster recovery planning.

Ukraine's swift response to the Zelensky deepfake demonstrates some of the top priorities in responding to disinformation from AI and can serve as an exemplar for governments, organizations, and individuals alike. First, the response was planned. Ukraine was prepared for potential Russian disinformation and had developed a playbook in advance for responding to deception in near-real-time. Information security incident response plans are required in most midsize to large organizations so that IT and other staff can respond quickly and effectively to cyber breaches, ransomware attacks, and similar events. Planning in advance for disinformation incidents against government agencies or officials, businesses, and individuals may be seen as a common compliance obligation in a similar fashion in the foreseeable future.

Further, the response to the alleged Russian deepfake video was immediate and took advantage of both traditional and online media. Within minutes of the fake video's appearance on television, President Zelensky posted a personal response via Facebook video repudiating the video's message and discrediting the deepfake (Simonite, 2022). News sources cited the speed and authenticity of the response as critical in dealing with the deepfake video and dispelling rumors before they could take root.

The video response was a live recording of the President, personalized and authentic to the situation, and it directly refuted the deepfake video's message and claim. Media outlets covered the incident, along with Zelensky's response. This could be applied to any organization's disinformation playbook by preparing public relations, social media, marketing, or similar staff to capture and disseminate authentic videos of the CEO or other targeted individuals addressing a deepfake video or false, AI-generated disinformation head-on.

In the Zelensky case, Facebook and YouTube eventually deleted uploads of the deepfake video, as deceptive or manipulated media is a violation of their terms of service, but it was Zelensky's quick, thorough, personal, and authentic response, combined with his public relations team's swift distribution of that message through both social media and traditional media outlets that neutralized the potentially harmful situation before it had a chance to be disseminated widely (Allen et al., 2023).

Developing a Disinformation Incident Response Playbook

Based on the Zelensky case study and preceding literature review, below are the authors' recommendations for governments, organizations, and individuals as they develop playbooks for responding to AI-generated or AI-assisted disinformation in near-real-time:

1. **Be prepared**. Develop a playbook or have a plan in advance for responding to disinformation events, just like your organization is likely required to do for cybersecurity incidents. Larger organizations should include handling disinformation via social media in their table-top preparedness exercises alongside natural disasters and cyber incidents for business continuity and disaster recovery scenarios.

2. **Respond quickly, personally, and authentically**, preferably with live video and audio of the subject of the disinformation. Zelensky's immediate response, in the form of a personalized smartphone video that was distributed first on Facebook and then through news outlets, was key to halting Russia's alleged disinformation operation. If a CEO, world leader, or individual is the subject of a deepfake, digital voice impersonation, fake images, misleading AI-generated online news articles, or other disinformation, that person must be ready to respond in near-real-time to remove any momentum from the fake media before it spreads beyond containment.

3. **Use traditional media and online, social media** to refute disinformation across all platforms. Beginning with the platform where the original disinformation was shared, which was Facebook in the Zelensky case, was a crucial step in Ukraine's response, but they didn't stop there. Zelensky's team quickly called in the aid of the nation's press and then world news organizations in curtailing the spread of the falsified deepfake video. Depending on the situation and the severity, posting a video to multiple social media platforms would be followed by calling a press conference or reaching out to local and national news outlets. In most cases, multiple forms of media should be engaged as part of an organization's playbook when fighting information warfare.

4. **Follow through** with news outlets and online sites to remove manipulated and deceptive media to ensure that it isn't re-disseminated later. In the Zelensky case, Facebook and YouTube removed the altered videos within hours, and news organizations superimposed the word "DEEPFAKE" on videos and still-frames featuring the manipulated media to help ensure it would not be reposted and misinterpreted as authentic. Sites like Facebook, X (formerly Twitter), YouTube, and other social media giants may take more time to remove posted videos or audio recordings, but they are a valuable part of the process. And include instructions in your organization's disinformation incident response playbook for your own public relations team to indelibly mark all falsified video, audio, and images as "DEEPFAKE", "FALSE INFORMATION", or "AI-GENERATED" so that they will be readily identifiable as fake, even when reposted or shared out-of-context.

5. **Train employees** to be suspicious of and verify not only email and text messages, but also phone and video calls. In every social engineering attack, awareness is a key factor. Social engineering awareness is probably already a part of your organization's cybersecurity incident response plan, but highlight the threat of deepfakes and AI-generated disinformation so your team can recognize and respond to attacks as quickly and effectively as the Ukrainian leadership team. Make your employees and top-level executives aware of the danger of deepfake voice and video calls, as well as AI-generated emails, text messages, web pages, and news articles. Educate and

empower employees to verify all requests before complying.

Testing the Playbook

How Organizations Can Respond to AI-Generated Fake News Postings

Let's apply the recommendations in the preceding section to the case of disinformation via massive fake news postings like the one about China's takeover of Taiwan presented earlier. First, business leaders and governments of both China and Taiwan should be prepared for disinformation events such as this one. Communications professionals in each government would need to develop a rapid repudiation message, using live video of both television personalities and government officials, preferably live in recognizable, public spaces in Taiwan, showing that there was no actual invasion. Such messages would need to be posted online and broadly disseminated via traditional news and media outlets. For Taiwan, the goal would be to prevent disruption of the economy and trade with peer nations. For China, it would be equally important to prevent international sentiment from turning against China and avoid sanctions, as well as other negative outcomes.

Responding to False Generative AI Images, Video, and Audio

Politicians, CEOs, organizations, and governments must consider the future need to counter disinformation in their planning exercises. The false images of the Trump arrest were noted by the original poster to be fictitious, just as the image of Clinton and Trump kissing was presented as AI-generated by the authors of this manuscript. But if similar, or even more damaging, images, video, and/or audio were released by a rival, an adversary, malicious government, criminals, or terrorists, both candidates and leaders alike would need to be prepared to respond immediately, personally, and authentically via live video to address the specific disinformation being spread.

The Ultimate Challenge: Responding to Real-Time Deepfakes and Disinformation

Experts across industries have been predicting for years that technology would advance to the point that real-time deepfake video, audio and similar disinformation using AI would be possible, and at the DEFCON hacking conference late last year, those fears were found to have come true (Zror, 2023). Security researcher Gal Zror demonstrated a combination of already-existing technologies chained together to replace his face and voice in real-time with the founder of DEFCON, Jeff Moss, with only a slight delay similar to what we have come to expect from long-distance web conferences or satellite correspondents' video during a live news broadcast. Zror jokingly misinformed the crowd that the DEFCON conference was canceled (Zror, 2023).

Responding to a malicious actor who has stolen both a leader's face and voice, and who can both speak with news agencies and post live videos to social media and online sources, may be the pinnacle of dispelling disinformation. In addition to all the components discussed above in forming a playbook for responding to disinformation, maintaining a personal relationship with multiple media outlets and personalities, or at least with a well-connected public relations firm, would help mitigate the damage such an actor could inflict. It could very well be a near-term need for an affected politician or business leader to personally call a news anchor on their mobile phone to show that they're the real person—but a highly resourced malicious actor could easily spoof the phone number of the celebrity or leader they're impersonating.

All elements of the playbook would need to be deployed quickly to be able to respond to publicly posted videos, but the final playbook recommendation, training our employees to be aware of the threat of deepfake social engineering, would be crucial to stopping targeted fraud and theft like this.

Beyond our own organizations, educating members of the media, and the general public, as well as ourselves, our friends, and family, of the present danger posed by AI-enhanced impersonators must become part of the strategy for dispelling disinformation before it spreads. The old-fashioned journalistic tenet of verifying a source and the information provided by that source has become both more difficult and more vital than ever. But as traditional media sources fade into the background amidst the clamor of misinformation inadvertently shared and re-posted by our own friends and online connections, the prospect of cutting off disinformation before it spreads becomes a matter of incident response and business continuity planning.

Conclusions

While the world has not yet experienced, as of the time of this writing and to the best of our knowledge, a large-scale information operation leveraging the full extent of deepfake misinformation, or fake news, audio, video, or images produced by generative AI, the near-term prospect has been demonstrated and is a current threat to leaders of nations, for-profit and non-profit organizations, and private individuals. In the not-too-distant future, a video teleconference with your CEO asking for an urgent wire transfer, a phone call with your nephew asking for money, or a public address by the leader of your nation claiming to have just stepped down from power may be realistic enough to fool the majority, or at least enough to have the intended effect of defrauding, distracting, disrupting, or worse.

This research extends and develops the recommendations of previous authors through the lens of the case study on Ukraine's response to alleged Russian disinformation operations against Ukrainian President Volodymyr Zelensky near the outset of the Russian invasion of Ukraine. The goal of this research is to encourage organizations and leaders to prepare a disinformation incident response playbook to respond to potentially damaging AI-generated content. This playbook approach advocates for targeted organizations and individuals to respond in real-time, using personalized, authentic messaging to dispel fake video, audio, or images. Furthermore, the target of a disinformation campaign should use both traditional media outlets and online/social media platforms together in repudiating misleading video/audio/images/text, and the affected individual or organization must follow up with online platforms after the fact to ensure that manipulated or AI-generated false information is taken down to avoid further redistribution.

Ultimately, citizens, investors, and leaders will have to rely either upon more advanced detection systems (algorithms driven by more AI) or upon better education for journalists, organizations, ourselves, and the public at large in verifying sources and information before sharing. Organizations and governments can plan to successfully mitigate the reputational damage and real-world destruction that can be wrought by ChatGPT, generative AI in general, and deepfake-manipulated media beginning with the playbook uncovered in this research.

In the meantime, we are left to contend with the fact that we simply cannot believe what we see or hear in audio, video, images, or text until it is checked, rechecked, and verified—something the intelligence community has known for decades, but now at a scale and velocity that can impact economic and global stability faster than a news cycle. But by developing a playbook for dealing with AI-generated or AI-enhanced disinformation ahead of time, we can enhance the effectiveness of our incident response and improve the odds of our businesses' or governments' survival from such foreseeable, tangible, and near-term threats.

References

Allen, C., Payne, B.R., Abegaz, T.T., Robertson, C.L. (2023). What You See Is Not What You Know: Studying Deception in Deepfake Video Manipulation. *Journal of Cybersecurity Research, Education, and Practice (JCERP)*, 2024 (1), Article 1, 7 pp. October 2023. ISSN 2472-2707. https://digitalcommons.kennesaw.edu/jcerp/vol2024/iss1/1/

Australian Strategic Policy Institute (ASPI). (2023). *ASPI's Critical Technology Tracker: The global race for future power* (Policy Brief Report No. 69/2023). ISSN 2209-9670. Retrieved from https://www.aspi.org.au/report/critical-technology-tracker

Biniyaz, J. (2023, May 17). Generative AI: Posing Risk of Criminal Abuse. *Readwrite.* Retrieved from https://readwrite.com/generative-ai-posing-risk-of-criminal-abuse/.

Bipartisan Policy Center (2020). *Artificial Intelligence and National Security.* June 2020. Retrieved from https://bipartisanpolicy.org/report/ai-national-security/

Blais, J. R., & Jungdahl, A. M. (2019). Artificial Intelligence in a Human Intelligence World. *American Intelligence Journal*, *36*(1), 108–113.

Clifford, C. (2019, March 26). Bill Gates: A.I. is like nuclear energy — 'both promising and dangerous'. *CNBC.* Retrieved from https://www.cnbc.com/2019/03/26/bill-gates-artificial-intelligence-both-promising-and-dangerous.html

Clifford, C. (2018, February 21). Top A.I. experts warn of a 'Black Mirror'-esque future with swarms of micro-drones and autonomous weapons. *CNBC.* Retrieved from https://www.cnbc.com/2018/02/21/openai-oxford-and-cambridge-ai-experts-warn-of-autonomous-weapons.html

Cox, J. (2023, April 13). A Computer Generated Swatting Service Is Causing Havoc Across America. *Vice.com*. Retrieved from https://www.vice.com/en/article/k7z8be/torswats-computer-generated-ai-voice-swatting

Defense Advanced Research Projects Agency (DARPA). (2018, September 7). *DARPA Announces $2 Billion Campaign to Develop Next Wave of AI Technologies*. Retrieved from https://www.darpa.mil/news-events/2018-09-07

Di Placido, D. (2023, March 22). AI-Generated Images of Donald Trump Getting Arrested Foreshadow a Flood of Memes, Fake News. *Forbes.* Retrieved from https://www.forbes.com/sites/danidiplacido/2023/03/22/ai-generated-images-of-donald-trump-getting-arrested-foreshadow-a-flood-of-memes-fake-news/

Farah, H. (2023, August 2). Humans can detect deepfake speech only 73% of the time, study finds. *The Guardian.* Retrieved from https://www.theguardian.com/technology/2023/aug/02/humans-can-detect-deepfake-speech-only-73-of-the-time-study-finds

Faunt, R. A., & Gentile, P. D. (2019). Artificial Intelligence within the Intelligence Community: The Need to Retain the Human Dimension. *American Intelligence Journal*, *36*(2), 48–53.

Green, J. I. A. S. W. B. (2024). Unraveling the Dynamics and Impacts of Financial Sabotage: A Comprehensive Analysis. *Researchgate.* Retrieved from https://www.researchgate.net/ profile/ Jamell-Samuels/publication/378653936_Maternal_Haplogroups_and_ Performance_in_Long-Distance_Running_A_Comprehensive_Analysis/

Isik, O., Joshi, A., & Goutas, L. (2024, May 31). 4 Types of Gen AI Risk and How to Mitigate Them. *Harvard Business Review.* Retrieved from https://hbr.org/2024/05/4-types-of-gen-ai-risk-and-how-to-mitigate-them

Kareem, T. A. (2024). Impact of Chat GPT on Human Communication and Social Interaction. *International Journal of Engineering and Modern Technology*, *10* (8), pp. 30-50. Retrieved from https://iiardjournals.org/

Liu, M. -Y., Huang, X., Yu, J., Wang, T.-C, & Mallya, A. (2021). Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications. In *Proceedings of the IEEE, 109*(5), pp. 839-862.

Lu, D. (2023, March 31). Misinformation, mistakes and the Pope in a puffer: what rapidly evolving AI can – and can't – do. *The Guardian.* Retrieved from https://www.theguardian.com/ technology/2023/apr/01/misinformation-mistakes-and-the-pope-in-a-puffer-what-rapidly-evolving-ai-can-and-cant-do

Kahn, J. (2023). Fake News 2.0: The Election Threat from A.I. *Fortune*. Retrieved from https://fortune.com/2023/04/08/ai-chatgpt-dalle-voice-cloning-2024-us-presidential-election-misinformation/

Mai, K. T., Bray, S., Davies, T., Griffin, L. D. (2023, August 2). Warning: Humans cannot reliably detect speech deepfakes. *PLoS ONE 18*(8): e0285333. Retrieved from https://doi.org/10.1371/journal.pone.0285333

Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof, 23*(3), pp. 255-262.

Maruf, R. (2023, May 28). Lawyer apologizes for fake court citations from ChatGPT. CNN. Retrieved from https://edition.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers/index.html

Martinovic, G., Bandur, K. M., & Tusevski, B. (2024, April). The Impact of Artificial Intelligence on Organizational Structure Trends Analysis and Implications. In *Economic and Social Development (Book of Proceedings), 110th International Scientific Conference on Economic and Social* (Vol. 5, p. 149).

Milligan, S. (2023, May 23). White House Pledges 'Road Map' for Managing Artificial Intelligence. *U.S. News.* Retrieved from https://www.usnews.com/news/politics/articles/2023-05-23/white-house-pledges-road-map-for-managing-artificial-intelligence

Moran, C.R., Burton, J., Christou, G. (2023). The US Intelligence Community, Global Security, and AI: From Secret Intelligence to Smart Spying. *Journal of Global Security Studies*, *8*(2), 2023.

Rana, M. S., Nobi, M.N., Murali, B., & Sung, A.H. (2022). *Deepfake Detection: A Systematic Literature Review. IEEE Access 10*, pp. 25494-25513. Retrieved from https://ieeexplore.ieee.org/abstract/document/9721302

Richterova, D., Grossfeld, E., Long, M., & Bury, P. (2024). Russian Sabotage in the Gig-Economy Era. *The RUSI Journal*, *169*(5), pp. 10–21. Retrieved from https://doi.org/10.1080/03071847.2024.2401232

Shapiro, C. (2022, October 4). The Intelligence Community Is Developing New Uses for AI. *FedTech.* Retrieved from https://fedtechmagazine.com/article/2022/10/intelligence-community-developing-new-uses-ai-perfcon

Simonite, T. (2022, March 17). A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be. *Wired*. Retrieved from https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/

Stupp, C. (2019, August 30). Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. *The Wall Street Journal.* Retrieved from https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

The Guardian (2022, March 19). Deepfakes v pre-bunking: is Russia losing the infowar? Retrieved from https://www.theguardian.com/world/2022/mar/19/russia-ukraine-infowar-deepfakes

The Guardian (2019, March 29). Conmen made €8m by impersonating French minister - Israeli police. Retrieved from https://www.theguardian.com/world/2019/mar/28/conmen-made-8m-by-impersonating-french-minister-israeli-police

The White House. (2022). *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*. Retrieved from https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Tucker, P. (2020, January 27). Spies Like AI: The Future of Artificial Intelligence for the US Intelligence Community. *Defense One*. Retrieved from https://www.defenseone.com/technology/2020/01/spies-ai-future-artificial-intelligence-us-intelligence-community/162673/

Urman, A., & Makhortykh, M. (2023, September 8). The Silence of the LLMs: Cross-Lingual Analysis of Political Bias and False Information Prevalence in ChatGPT, Google Bard, and Bing Chat. *OSFPREPRINTS*. Retrieved from https://doi.org/10.31219/osf.io/q9v8f

United Kingdom Government (n.d.). *Pioneering a new National Security. The Ethics of Artificial Intelligence.* Retrieved from https://www.gchq.gov.uk/artificial-intelligence/accessible-version.html

Zror, G. (2023). Look Ma, I'm the CEO: Real-Time Video and Audio Deepfakes. *DEFCON 31*. Retrieved from https://media.defcon.org/DEF%20CON%2031/DEF%20CON%2031 presentations/

QRBD

Quarterly Review of Business Disciplines

February 2025

Volume 11
Number 3/4

A Journal of International Academy of Business Disciplines
Sponsored by University of North Florida
ISSN 2334-0169 (print)
ISSN 2329-5163 (online)