Malinformation, Deepfakes, and Cyber Warfare: Ethical and Anticipated Ethical Issues

Noah Donnelly[1] and Richard Wilson[1,2]
[1]Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA
[2]Department of Philosophy, Towson University, Baltimore, Maryland, USA

ndonnel1@students.towson.edu
wilson@towson.edu

Abstract: Malinformation and deepfakes are closely connected in cyber warfare because both involve the use of *true* or *realistic-looking* information in harmful ways. Malinformation is *genuine (truthful) information used maliciously* to cause harm. Unlike disinformation (false) or misinformation (false but unintentional), malinformation is based on truth, but truth that has now been weaponized. Examples include, leaking private emails, exposing sensitive military data, or releasing authentic documents but out of context. Deepfakes can be AI-generated synthetic media (video, audio, images) that convincingly mimic real people or events. They can be employed to fabricate speeches, fake evidence, or impersonate leaders. Malinformation and Deepfakes Intertwine in several ways. They involve authenticity exploitation. Deepfakes often mix real content with fabricated elements. When genuine material is combined with manipulated media, it becomes malinformation — because the *truthful parts* lend credibility to the fake narrative. *Example:* Real leaked emails paired with a deepfake video of a politician "admitting" corruption. They promote context manipulation. Malinformation thrives on taking real information out of context. Deepfakes amplify this by visually or audibly presenting "evidence" that seems authentic. *Example:* A real battlefield video edited with deepfake audio to suggest atrocities that didn't occur. They have a psychological impact. Malinformation already erodes trust by exposing sensitive truths. Deepfakes magnify this by making it harder to distinguish between genuine leaks and fabricated ones. *Example:* Soldiers' real identities leaked (malinformation) alongside deep-fake videos of them committing crimes. Cyber warfare Strategy. Both are used to destabilize societies, discredit leaders, and manipulate public opinion. Deepfakes transform malinformation into a more persuasive weapon by adding *visual proof*. Analysts' warn that the "malinformation–deepfake nexus" is one of the most dangerous aspects of modern cyber warfare. Malinformation and deepfakes become tools of cyber warfare by bringing truth and illusion together, making it extremely difficult for societies to defend against manipulation. Malinformation and deepfakes intersect with cyber warfare because both weaponize elements of truth and realism to cause harm, erode trust, and destabilize societies. Deepfakes often transform malinformation into a more persuasive and damaging tool. This analysis will identify the ethical and anticipated ethical issues with the use Malinformation and Deep Fakes in Cyber Warfare.

Keywords: Malinformation, Deepfakes, Generative adversarial networks (GANs), Epistemic backstop, Liar's dividend, C2PA, Information disorder

1. Introduction

Within the conflicts of the modern era, the battlefield has grown past kinetic destruction of infrastructure to an additional focus on the cognitive alteration and destruction of reality itself. Wardle and Derakshan (2017) categorize this as "Information Disorder," which is characterized by three unique traits: Misinformation (unintentional lies), Disinformation (intentional lies), and Malinformation (weaponized truth). Cyber warfare traditionally focuses on the disruption of networks, where the new frontier is on the corruption of data. Two technologies have been the primary offenders in terms of this corruption: Malinformation, which uses breached private data to destroy people and institutions reputations, and Deepfakes, which uses Generative AI to create fake evidence.

The combination of Deepfakes and Malinformation creates a "Liar's Dividend," an idea introduced by Chesney and Citron (2019). The Liar's Dividend posits that when high-quality fake media becomes more prevalent, rational actors will begin to discount all media, media that includes organic evidence of corruption or war crimes. A state actor now has the capability of leaking real compromising files (Malinformation) and of combining them with deep-fake videos that create a fake but incriminating context. The point of these actions is not just to fool people, but to pollute the information pool so much that the targeted population begins to disengage from politics entirely, this tactic is known as "censorship through noise." This paper analyzes how the technical mechanisms behind Generative Adversarial Networks (GANs) violate the ethical "Right to Truth," and applies Anticipatory Ethics to the rampant, unchecked proliferation of open-source synthetic media technologies. Through the cases of the Ukraine conflict, the Gabon coup attempt, and the Macron Leaks, we show that the target of modern information warfare operations is the human mind itself.

2. Technical Issues

Synthetic media is being weaponized, and this weaponization is being driven by the rapid commoditization of deep learning models. Early deepfakes relied on basic autoendcoders, neural networks capable of learning data coding's in an unsupervised way, the current threat is defined by three, very advanced technical artifacts: Latent Diffusion Models, Transformer-based Audio Synthesis, and Adversarial Counter-Forensics (Mirsky & Lee, 2021).

2.1 Latent Diffusion Models

General Adversarial Networks (GANs) originally sparked the deepfake era by the "minimaxing" game between Generator and Discriminator. The new method is Latent Diffusion Models (LDMs). GANs map noise directly to an image, LDMs are far more advanced, they operate by iteratively denoising a random Gaussian noise field based off text or semantic inputs (Rombach et al., 2022). This allows for more stability and a higher resolution then GANs have, which typically suffer from "mode collapse," where the generator would output an image identical to the one it was provided. In the context of cyber warfare, LDMs represent force multipliers, the feature of "Inpainting" (ability to edit specific regions of a real image) and "Outpainting" (ability to insert compromising elements like nudity, illicit substances), all done with perfect coherence. LDMs lower the technical barrier for high-fidelity forgeries. LDMs simply do not need the complex hyperparameter tuning that GANs require (Mirsky & Lee, 2021).

2.2 Zero-Shot Audio Synthesis and Transformers

Possibly the most frightening evolution is the rise of Zero-Shot Text-to-Speech (TTS). Previous models like Tacotron 2 needed hours of a targets voice to train a clone of a voice. Modern TTS's are built on Transformer models (similar to GPT-4) which treat audio synthesis as a language modeling task. By turning audio waveforms into acoustic tokens, models such as VALL-E can mimic the targets voice, prosody, and emotional tone in as little as three seconds of audio as a reference (Wang et al., 2023). Using this capability, attackers can simply intercept a phone call, train a clone, and disseminate falsities in a matter of minutes (Malinformation). The attacker would take a real, leaked phone call, use a Speech-to-Speech translator that retains the target's vocal signature and the correct background noise, which makes modern spectral analysis ineffective (Westerlund, 2019)

2.3 Adversarial Perturbation and Counter-Forensics

With the improvement of detection algorithms, attackers are adapting by deploying Adversarial Perturbations. This involves injecting a specific pattern of low-magnitude noise, invisible to humans, into the deepfake which mathematically blinds the detection algorithms (Carlini & Farid, 2020). An example of such tactics would be an attacker adding a gradient-based noise layer that would force a RestNet-50 detector to mark a fake video as "Real" with 99% certainty. This is the "Arms Race" that is dynamically unfolding on the digital stage; aggressors optimize their content to defeat the forensic tools that platforms like Facebook or Twitter employ. Passive detection is unreliable because the detector is now the target of the generative process.

2.4 Malinformation Infrastructure

The production of Malinformation is no longer a manual process, it is automated using Natural Language Processing (NLP) pipelines. When an Advanced Persistent Threat (APT) steals terabytes of data, attackers use Named Entity Recognition (NER) to automatically scan millions of files for sensitive keyworks like "confidential," "bank," "health," and "affair." The leaks are then indexed by the systems which now allow intelligence officers to query for compromising material related to specific targets. Now automated, it allows for leaks to be "Micro-Targeted" where subsets of private data are released to a target demographic on social media to maximize psychological impact. This is known as "Precision Propaganda" (Starbird et al., 2019).

3. Ethical Issues

Deploying these technologies in warfare raises important questions regarding the "Right to Truth."

3.1 Eroding the Epistemic Backstop

Philosopher Regina Rini (2020) argued that audiovisual recordings serve the purpose of an "Epistemic Backstop," which is a final, objective verifier of 'what is reality' that resolves disputes. When that verifier (video evidence) is undermined, truth itself becomes hard to ascertain and since the backstop is removed, society enters a state of "Epistemic Insecurity." This attacks the autonomy of the citizens themselves, which can be defined as an attack on a civilian population, which is an ethical violation of the laws of war. If citizens cannot determine if a video about a presidential candidate is real or not, they can no longer make a free and rational decision about

the reliability of the candidate. This causes the degradation of the democratic process and moves society towards tribal beliefs and not evidence-based good, spirited debate on important issues. Fallis (2020) expands upon this, arguing that the existence of the technology causes harm by creating "Technological Defeasibility," which is the action of disregarding true evidence believing it is fake. This allows authoritarian regimes to write off documented human rights abuses as "Western AI fabrications," protecting these regimes from accountability and the repercussions that come as the result of accountability.

3.2 Digital Violence and Contextual Integrity

"Contextual Integrity" proposed by Nissenbaum (2019) is an ethical framework violated by Malinformation. Malinformation violates the typical channels in which information flows, typically governed by context-specific norms (medical data staying with a doctor), is now taken out of context and weaponized by forcing private information out of its intended context, into the public realm. Even if the information is technically true, its release is intended to coerce or shame the target. In cyber warfare, this is targeting the "Human Layer" of the Open Systems Interconnection (OSI) model (Zimmerman, 1983). By exposing the private lives of important individuals like diplomats, politicians, and soldiers, the attackers aim at causing psychological trauma that can destroy their adversary's motivation to fight. This is a violation of the Just War principle of Proportionality (May, 2007), as the long-term effects upon the targets psyche represent damage to non-combatants (in a reputational sense), which far exceeds the military advantage gained through kinetic warfare.

4. Case Studies

Cyber criminals leverage these attacks to maximize economic chaos and instigate liquidity crises (Kenton, 2023). The once theoretical threat of deepfake-malinformation has already materialized in global conflicts, altering the course of history. These cases showcase how state and non-state actors use the "Liar's Dividend" and "Tainting" strategies to bypass traditional defense systems against deepfakes.

4.1 Zelenskyy Fake Surrender Video (2022)

In March of 2022, a hacker group compromised the systems of the Ukrainian news station called *Ukraine 24* by hijacking the live news feed to stream a deep faked video of President Volodymyr Zelenskyy. In this fake video, Zelenskyy stood behind a podium and ordered all Ukrainian troops to lay down their arms and surrender themselves to the Russian forces. Even though the quality of the video was low, it featured a static body and mismatched pixelation around Zelenskyy's neck, nevertheless the operation showed the concept of a "Deception Strike" via information warfare. The hackers didn't just rely on social media, they compromised a "root of trust," that being the national news broadcast, to attempt to lend credibility to the fake video. Atlantic Council analysts noted that while this attempt failed due to rapid debunking, it was considered a "Proof of Concept" for future operations. If such a video was released during a time of blackout or nuclear warfare, soldiers on the front lines might obey the order before it could be fact checked. This shows the dangers of deepfakes in "Time-Critical Disinformation" operations related to kinetic warfare (Atlantic Council, 2022).

4.2 Gabon Coup Attempt (2018)

In late 2018, the president of Gabon, Ali Bongo, had not been publicly seen for months due to a stroke, which made a power vacuum fueled by rumors of his death. To create an image of stability, the government released a "New Year's Address" video. The video was uncanny, observers noticed the president's eyes did not blink for long periods of time and his facial movements appeared to be like someone wearing a mask. The political opposition instantly claimed that the video was a deepfake, and proof that the president was either dead or incapacitated. Claiming that this "fake" video was authentic meant that the government had lost all legitimacy, As a result Lieutenant Kelly Ondo Obiang launched a military coup on January 7, 2019. This case is an example of the "Liar's Dividend": the existence of deepfake technology alone had allowed the rebels to weaponize speculation, transforming a low-quality video to support rumor and using the video as a justification for a violent military coup. This gave evidence that deepfakes do not need to be high quality to be lethal, they only need to be plausible enough to confirm an existing bias (Breland, 2019).

4.3 Macron Leaks (2017)

Hackers linked to the Russian GRU released 9 gigabytes of emails that were stolen from Emmanuel Macron's campaign in an operation called #MacronLeaks, two days before the French presidential election. This was a professional "Malinformation" operation that used a technique called "Tainting." The dump was primarily boring, authentic administrative emails, but the hackers forged a small number of documents suggesting that

Macron had offshore bank accounts containing illicit funds. The release was timed for Friday night before the election, triggering the mandated 44-hour media blackout in France. This timing was planned so that journalists could not verify the documents before the voting began, forcing the French public to rely on social media rumors. The operation only failed because the campaign had flooded their own servers with a “honeypot” of fake data which confused the attackers. Nevertheless, it established a doctrine of “Hack-and-Forge,” where truth is used to deliver lies (Jeangène Vilmer, 2019).

4.4 Pentagon Explosion Hoax (2023)

In May of 2023, a verified Twitter account by the name of “Bloomberg Feed” (a fake account trying to pass themselves off as a legitimate news source) posted an AI-generated image of smoke clouds rising from the Pentagon and claiming that a massive explosion had occurred. The image was artificially boosted by the Russian state media *Russia Today (RT)* and thousands of bots. In less than an hour the S&P 500 index dropped by 0.3%, wiping out $500 billion in market capitalization. This was a realization of the vulnerability in the High-Frequency Trading (HFT) algorithms. These trading bots scan social media for breaking news keywords and execute trades in milliseconds, far faster than any form of human verification is possible. The hoax served as a probe of the stock market’s reaction time, which suggests that deepfakes could be timed to coincide with real world kinetic activities.

5. Anticipatory Ethics

Applying the "Moral Responsibility for Computing Artifacts" framework by Miller et al., 2011 to the developers who engineered Generative AI systems to engineer deepfakes reveals an extreme failure of foresight. The rules are aimed at developing normative rules for people who design, develop, deploy, evaluate or use computing artifacts. The invention of Generative Adversarial Networks (GANs) is not the primary ethical failure here, it is the reckless increase in their capabilities with no countermeasures, and the possibility of their proliferation for anyone who wishes to use them.

5.1 Rule 1 and Dual-Use Dilemma

According to Rule 1, “The people who design, develop, or deploy a computing artifact are morally responsible for that artifact, and for the foreseeable effects of that artifact.” This principle states that responsibility cannot be placed upon the consumer. Although AI that was designed for dubbing Hollywood movies it can now be leveraged to commit acts of fraud, this is the “Dual-Use” dilemma. Brundage et al. (2018) argues that when the cost of generating fake media drops, it is foreseeable that bad actors will use these AI tools for “automated spear-phishing” (AI targeting specific people with personalized prompts) and "high-efficiency disinformation" (AI used to spread high-speed, low-cost disinformation) (Bipartisan Policy Center, 2023). Releasing open source deepfake repositories (DeepFaceLab) with zero watermarks or user restrictions violates Rule 1 of Anticipatory Ethics. Developers prioritizing science over safety is no new phenomena, they yet again ignored the foreseeable risk that state actors can weaponize their libraries (Code Repositories) to take advantage of democratic institutions. The developers of this library entirely failed the “foreseeability effect” allowing their tools to be used to destabilize the global information system.

5.2 The Sociotechnical Imperative and the Collingridge Dilemma

Rule 4 requires that “The people who design, develop, or deploy a computing artifact must ensure that the artifact is designed in such a way that it considers the sociotechnical system in which it is embedded.” Developers should now be required to foresee how their products will be used in a low-trust political environment. “Epistemic Insecurity,” which is a low trust environment where citizens struggle to verify what is true and what is not, is the current state of the information system. With deep-fake forgeries being undetectable, the already fragile system begins to break further. This is exemplified perfectly in the Collingridge Dilemma (Collingridge, 1980) which when applied would state: early on (2014-2017) the technology was able to be controlled and combatted, but they did not see the need to do so. Currently (post 2023), the impact of the technology has been more completely realized, and the developers have lost the ability to control their computing artifact due to the great number of open-source models (Brey, 2012). Anticipatory Ethics shows the need for Generative AI models to be “Red Teamed” for sociotechnical risks, such as testing their ability to mass persuade and defame, and this must be done pre-deployment and before the technology is introduced.

5.3 Rule 3: Responsibility to Retract

Rule 3 requires “if a computing artifact is found to be harmful, the developer has a moral responsibility to retract, modify, or mitigate the harm, even after deployment.” A duty of un-ringing the bell is now created for

developers. While they cannot un-invent GANs, the major AI labs have the moral duty to shift research from generation to detection. Vaccari and Chadwick (2020) argue the main harm caused by deepfakes is not fooling the public, it is the uncertainty it causes, leading the public to doubt authentic news sources. The developers who profited by creating generative tools have a duty to compensate the public by distributing advanced detection tools that are needed to restore the epistemic baseline which would allow for distinguishing what is real from what is fake. Yet they continue to distribute more powerful generative models without providing the detection software to combat them. This is a "Moral Hazard" where the developers get to profit from the suffering of society, while eventually the people pay for what the developers do.

6. Recommendations

To eliminate the threats of deepfakes being used for malinformation, we must move beyond passive verification methods like provenance and deploy Active Defense measures that disrupt the process of generating content and throttle the spread of falsified information.

6.1 Adversarial Data Poisoning

A new first line of defense is standardizing Adversarial Data Poisoning tools, this concept relying on the "Fawkes" framework made by Shan et al. (2020). This is far better than platforms trying to make their platforms immune to this via detection by disrupting the generative process itself. Deepfake models need incredibly large datasets of high-quality images to create what appears to be accurate content, Fawkes would disrupt this by injecting "cloaks," which are imperceivable, pixel level perturbations into user photos before they are posted to the internet. Cloaking algorithms solve an optimization problem, it decreases the perceptual difference (L2 distance) between the original and altered image, and it increases the distance in the "feature space" of normal standard recognition models (FaceNet or VGGFace). When a state-actor collects these "cloaked" images to train a recognition model, the model locks onto the invisible perturbations as opposed to the subjects real facial features, this disrupts the feature extractor making the images look distinct enough from the target so that the fake is no longer believable. This is called poisoning the well of training data, a small amount of faulty data can disrupt prediction-based models (like Generative AI) even if they have an overwhelming amount of authentic, clean data.

6.2 Latent Space Watermarking

To solve the attribution problems for synthetic media used in "Tainting" operations, we recommend that developers implement Latent Space Watermarking. Unlike classical metadata, which can be stripped or pixel-space watermarks that can be cropped, Wen et al. (2023) propose "Tree-Ring Watermarks." These new proposed watermarks entangle themselves within the generative process itself. Diffusion models make images by iteratively denoising a random Gaussian noise vector. A Tree-Ring watermark would change the initial noise vector through embedding a pattern in its Fourier transform (frequency domain). Since the diffusion process, mathematically, is continuous, the frequency pattern persists through the rest of the denoising chain and would be recognizable in the final image's structure. This watermark effectively hedges against someone trying to transform an image geometrically (rotating, cropping, JPEG compression). If a bad actor were to use commercial API like DALL-E or Midjourney to create malinformation, the watermark would allow forensic analysts to undo the diffusion and retrieve the "Key" that was used to generate the image. This acts as a mathematical proof of origin; the platforms can then find who is to blame for the attack and revoke the API access instantly.

6.3 Frequency Domain Forensic and Spectral Analysis

Deepfake detectors currently suffer from "overfitting," which means they learn to spot specific visual cues (mismatched earrings, lack of blinking, or stiff facial gestures), which current models correct in their diffusion process. Developers cannot keep making detectors for levels of technology people do not use anymore, instead it needs to be future proofed. To "Future-Proof," we must move over to Frequency Domain Forensics. (Frank et al. 2020). This shows that while GANs produce spatially perfect images, they leave massive and distinct artifacts in their frequency. Generative models use "Up-Sampling" layers (like Transposed Convolutions) to scale low resolution vectors into a high-resolution image. Up-sampling inserts grid-like patterns into the image's spectrum, which can be made visible by applying a Discrete Cosine Transform (DCT) or Fourier Transform. Spectral anomalies such as these are inherent to the Generative models' architecture and would be incredibly hard for them to unlearn. We recommend Adversarial Training pipelines that can target these grid patterns to be implemented into current defense systems. If you can train a detector to ignore what is happening in the pixel

space and look at the spectrum of the image, defenders can now accurately detect deepfakes from even unknown models, which is a proactive defense.

6.4 Algorithmic Circuit Breakers

Technical verification can take hours of time; this always gives the attackers the first move advantage. Platforms need to implement Algorithmic Circuit Breakers. Using the "Empathic Media" framework by Bakir and McStay (2018), this recommendation can create a speed bump against the spread of viral disinformation, which would punish velocity. The conditions to trigger this algorithmic circuit breaker would be an implemented "Velocity-Veracity Ratio," being installed, which would look at a piece of high stakes media as it begins to go viral, if the viral velocity exceeds the 99th percentile of organic growth, the algorithm will automatically trigger a "Circuit Breaker." Due to censorship concerns this would not automatically delete the content, but it would throttle the algorithms amplification of this material for a cooling off period (e.g., 60 minutes). Introducing this friction would stop the flood of falsehood from overwhelming public debate and would buy time for independent forensic analysts to catch up to the light-speed attack.

7. Future Work

Generative AI has begun its transition from image-only (unimodal) to multimodal architecture. The threat is going to move from "public disinformation" to "individualized reality distortion." Future research must think about this proactively and address three emerging issues that threaten the security of the internet.

7.1 Multimodal Chimeras and Consistent History Generation

The current deep-fake tools produce detectable content because they lack consistency (audio does not perfectly match the lip sync, or the document dates do not align with the video). Future work needs to address the anticipated rise of Multimodal Chimeras, these would be unified AI models that can generate a consistent, fabricated reality across every data type of media simultaneously. Horvitz (2022) warns about upcoming models that won't just generate a fake video, they will instantly generate a supporting chain of evidence, with a chain consisting of fake emails, calendar logs, and voice memos that align with the dates and times of the video. Research is urgently needed into "Cross-Modal Forensics" (checking to see if the lighting in the video matches the weather metadata in the logs). The goal of this research would to be able to detect these synchronized fabrications before they can be accepted as legal evidence.

7.2 Real-Time Biometric Injection

Liveness Detection systems are used in banking and high security access professions (Know Your Customer/KYC), this presents a serious technical threat. Future work needs to address Real-Time Deepfake Injection where attackers can take control of the virtual camera to feed a live, deepfake video into an authentication system. Moseley and Rhodes (2023) show that "Passive Liveness" checks, which look for blinding or changes in blood flow, are becoming increasingly vulnerable to diffusion-based video synthesizers that can accurately mimic human physiology. We predict a collapse in the effectiveness and use of remote identity verification and a requirement to return to "Hardware-Based Attestation," like physical keys to a door for accessing critical infrastructure.

7.3 Personalized Reality Distortion

Current cyber warfare focuses on large audiences (like in the Pentagon Hoax), future operations will likely use Personalized Micro-Targeting. By combining Large Language Models (LLMs) with voice cloning models, attackers will be able to perform spear-phishing at an enormous scale. Future research should model the effectiveness of "One-to-One" deepfakes, where an AI bot can call an employee using their bosses voice and talk to the employee in a fluid, dynamic conversation to authorize the transfer of funds to an offshore account. Hwang (2020) states that this commoditization of deception will make "security awareness training" useless, as fake media will be indistinguishable from reality, needing a purely technical defense, which cuts out the weakest link in cyber-security, the human one.

8. Conclusion

The combination of malinformation and deepfakes represents a major issue in cyber warfare, it establishes a point where the cost of generating fake, but convincingly real media drops to zero, while the cost of verification only increases with technological advance. Through our analysis of "minimax" game that is inherent to General Adversarial Networks and the Collingridge Dilemma that faces regulators, we conclude that the combination of

malinformation and Generative AI is not a small problem, but an existential threat to individual autonomy and sovereignty.

The case studies of the Gabon Coup and the Pentagon Hoax show that we are already living in the era of the "Liar's Dividend," where the possibility of deepfakes alone is enough to potentially destabilize governments and destroy economies. Schick (2020) argues that we are facing an "Infocalypse," which is a collapse of the shared epistemic framework needed for society to function properly. If ordinary citizens cannot discern between a real, leaked video of a war crime and a synthetic video of someone being shot by a militiaman, political accountability and discourse become impossible, and the residents of the parent state devolve into "Epistemic Nihilism."

By applying Anticipatory Ethics, we can recognize that this future is not set in stone, shaping it is a matter of engineering choices. One solution to the problems related to deepfakes is changing our defensive strategy to "Trust Architecture" (cryptographic) from "Truth Verification" (which is subjective). By installing Adversarial Data Poisoning to hedge image and video generation, Latent Space Watermarking to find the culprits, and Algorithmic Circuit Breakers to curb trending disinformation, misinformation, and malinformation, we can rebuild the "Epistemic Backstop" foundational to the internet. The integrity of our information must be treated with the same severity and urgency as a missile strike on a major city would be, a new social contract must be formed and enforced where a person's "Right to Reality" is an inalienable right.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: ChatGPT 5.1 for drafting and refinement. Human authors verified all content. Gemini 3.0 pro used for aiding in sourcing.

References

Atlantic Council. (2022). *Deepfakes in the War in Ukraine: A Threat Assessment*. Digital Forensic Research Lab (DFRLab).

Balis, C. (2020). The Future of the Truth: Deepfakes and the Threat to Democracy. *Harvard International Review*.

Bipartisan Policy Center. (2023, August 23). *Generative AI and disinformation*. https://bipartisanpolicy.org/article/generative-ai-and-disinformation/

Bond, S. (2023). Fake AI photo of an explosion at the Pentagon briefly creates confusion. *NPR*.

Breland, A. (2019). The Bizarre Story of the "Deepfake" That Might Have Sparked a Coup in Gabon. *Mother Jones*.

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *NanoEthics*, 6(1), 1–13.

Brundage, M., et al. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Future of Humanity Institute, University of Oxford.

Carlini, N., & Farid, H. (2020). Evading Deepfake Detectors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Chesney, R., & Citron, D. K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107, 1753.

Collingridge, D. (1980). *The Social Control of Technology*. St. Martin's Press.

Day, J. D. and H. Zimmermann, "The OSI reference model," in Proceedings of the IEEE, vol. 71, no. 12, pp. 1334-1340, Dec. 1983.

Fallis, D. (2020). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, 34, 623–643.

Frank, J., Eisenhofer, T., & Schönherr, L. (2020). Leveraging Frequency Analysis for Deep Fake Image Recognition. *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Goodfellow, I., et al. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*.

Horvitz, E. (2022). On the Horizon: Interactive and Compositional Deepfakes. *Proceedings of the 2022 ACM International Conference on Multimodal Interaction*, 1–10.

Hwang, T. (2020). *Deepfakes: A Grounded Threat Assessment*. Center for Security and Emerging Technology (CSET).

Jeangène Vilmer, J. B. (2019). The "Macron Leaks" Operation: A Study in Hybrid Warfare. *Journal of Cyber Policy*, 4(1), 42-64.

Kenton, W. (2023). *The Pentagon AI Hoax and the Vulnerability of Algorithmic Trading*. Investopedia Market Analysis.

Miller, K.W., et al. (2011). Moral Responsibility for Computing Artifacts: 'The Rules'. *IT Professional*, 13(3), 57–59.

Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, 54(1), 1–41.

Moseley, E., & Rhodes, M. (2023). The Threat of Deepfakes to Biometric Authentication and Liveness Detection. *Biometric Technology Today*, 2023(3), 5–9.

Nissenbaum, H. (2019). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosopher's Imprint*, 20(24), 1-16.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Schick, N. (2020). *Deepfakes: The Coming Infocalypse*. Grand Central Publishing.

Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., & Zhao, B. Y. (2020). Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. *Proceedings of the 29th USENIX Security Symposium*, 1589–1604.

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–26.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1).

Wang, C., Chen, S., Wu, Y., et al. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2301.02111*.

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, 9(11), 39–52.

Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy making*. Council of Europe.

Wen, Y., Kirchenbauer, J., Geiping, J., & Goldstein, T. (2023). Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. *Advances in Neural Information Processing Systems* (NeurIPS), 36, 68903–68923.

Wither, J. K. (2023). Weaponizing the Truth? The Danger of Malinformation in the Gray Zone. *Connections: The Quarterly Journal*, 22(2), 25-45.