Canadian Security Intelligence Service
Service canadien du renseignement de sécurité

The Evolution of Disinformation: A DEEPFAKE FUTURE

Canada

World Watch: Expert Notes series publication No. 2023-10-01

Aussi disponible en français sous le titre : *Évolution de la désinformation : un avenir « hypertruqué »*

This report is based on the views expressed during, and short papers contributed by speakers at, a workshop organised by the Canadian Security Intelligence Service as part of its Academic Outreach and Stakeholder Engagement (AOSE) and Analysis and Exploitation of Information Sources (AXIS) programs. Offered as a means to support ongoing discussion, **the report does not constitute an analytical document, nor does it represent any formal position of the organisations involved.** The workshop was conducted under the Chatham House rule; therefore no attributions are made and the identity of speakers and participants is not disclosed.

www.canada.ca

Published in October 2023
Printed in Canada

Cat. No. PS74-19/2023E-PDF
ISBN: 978-0-660-49645-0

The Evolution of Disinformation
A Deepfake Future

Highlights from an unclassified joint workshop of the
Academic Outreach and Stakeholder Engagement (AOSE)
and Analysis and Exploitation of Information Sources (AXIS)

24 May 2023, Ottawa

Table of Contents

The Workshop and its Objectives

On 24 May 2023, the Canadian Security Intelligence Service (CSIS) Academic Outreach and Stakeholder Engagement (AOSE) and Analysis and Exploitation of Information Sources (AXIS) programs co-hosted a workshop to examine the complex set of threats posed by deepfake disinformation technologies.

Held under the Chatham House rule, the workshop was designed around the work of eight leading experts from across the open-source research community, and benefited from the insights of security practitioners representing a range of domestic and international experiences. The papers presented at the event form the basis of this report. **The entirety of this report reflects the views of those independent experts, not those of CSIS.**

The CSIS Academic Outreach and Stakeholder Engagement program seeks to promote a dialogue between intelligence practitioners and leading specialists from a wide variety of disciplines working in universities, think-tanks, business and other research institutions. It may be that some of our interlocutors holds ideas or promote findings that conflict with the views and analysis of CSIS, but it is for this specific reason that there is value to engage in this kind of conversation.

Executive Summary

This report is based on the views expressed during, and short papers contributed by speakers at, a workshop organised by the Canadian Security Intelligence Service as part of its Academic Outreach and Stakeholder Engagement (AOSE) and Analysis and Exploitation of Information Sources (AXIS) programs. Offered as a means to support ongoing discussion, ***the report does not constitute an analytical document, nor does it represent any formal position of the organisations involved.*** *The workshop was conducted under the Chatham House rule; therefore no attributions are made and the identity of speakers and participants is not disclosed.*

The threats posed by disinformation to security and democracy have been assessed as a significant and ongoing, if not habitual, concern. Spurred by advancements in artificial intelligence (AI), deepfakes are viewed as a modern evolution of disinformation which poses new challenges for governments, individuals, and societies. Safeguarding the integrity of the information ecosystem is a fundamental priority not only for democracy, but also for society as a whole.

Technological Advancements and Prosocial Applications

Deepfakes, originally a portmanteau of deep learning and fake media, is now used more broadly to refer to any impersonating media created or edited by deep learning algorithms. Manipulated videos, images, audio/voice, and text created using generative AI techniques have quickly evolved to become increasingly accessible and realistic. In many ways, these advancements pose exciting opportunities.

- Developments in generative AI are significantly boosted by the availability of large-scale language and image generation models. Developments have focused on making the models more powerful, capable, and accessible while giving users more control over the style and content of the generated media through detailed text prompts.

- Deepfakes can be used for creating entertaining content, such as realistic face swaps or visual dubbing for film, TV, or video games, enabling new creative possibilities and bringing fictional characters to life. They can also be employed for visual effects and restoration purposes, recreating or enhancing scenes that are difficult or costly to produce practically. For example, deepfakes can be used to age or de-age actors or bring back deceased performers for a film or advertisement.

- As marketing and educational tools for training and simulations, deepfakes can enable realistic scenarios in fields like medicine, military training, or emergency response, allowing practitioners to practice skills in a safe and controlled environment.

Threats to Society and Security

As the capacity for generating media becomes more widely available and precise, the probability for misuse intensifies. Among the primary concerns with deepfakes is the potential for spreading disinformation and manipulating political discourse, leading to confusion, distrust, and social instability in democratic societies.

- Deepfakes raise serious privacy concerns as they can be used to create non-consensual explicit content by superimposing someone's face onto explicit material. This poses a threat to individuals' privacy and reputation while inflicting emotional distress. In addition, deepfakes present numerous legal and ethical challenges. They can infringe upon intellectual property rights and violate privacy laws.
- Deepfakes have the added potential to erode trust in visual media. As the technology becomes more sophisticated, it becomes increasingly challenging for people to distinguish between genuine and manipulated content, making it difficult to rely on video evidence and exacerbating the problem of disinformation.
- The widespread availability of deepfakes can have negative societal impacts, including cyberbullying, harassment, and potential for social unrest. Deepfakes can be weaponized to exploit or manipulate individuals, leading to reputational damage, psychological harm, or social divisions.

While deepfakes are more likely to advance already existing security threat-related activities rather than generating new concerns, it is important to recognize the potential risks associated with deepfakes and develop robust technological solutions, ethical guidelines, and legal frameworks to address these challenges and mitigate their negative consequences.

- Disinformation is a tool that has been used by state and non-state actors throughout history in their attempts to discredit and downplay democratic institutions, amplify conspiracies and radicalization, and encourage distrust of authority. Deepfakes facilitate the speed and effectiveness of these efforts, while expediting targeting of government/military personnel, phishing and social engineering, and mimicking of biometric data.

- Deepfakes can be used as a tool for creating noise to flood the intelligence collection space, causing distractions from true intelligence and/or distorting perceptions of human sources by creating artificially generated conversations, videos, or text. The increased reliance on open source intelligence (OSINT) makes deepfake information particularly impactful in the information ecosystem.

- Deepfakes can also be used to poison the data utilized for training of deep learning systems, intentionally compromising these systems with malicious information. For example, algorithms used to detect cyber-attacks could be compromised through data poisoning of the large-scale datasets on which they are trained.

- From a public safety and security perspective, deepfakes can be employed to commit fraud, engage in coercion and/or extortion, create fake evidence for criminal activities, or to impersonate and/or incriminate individuals in unlawful activities.

Outlook

Deepfakes are designed to deceive, and the human mind cannot consistently identify the outputs of sophisticated technologies. While tech giants have begun flagging deepfake content as disinformation, detection systems integrating both human and model predictions are of greater value. Governments have a role to play in facilitating the application of deepfake technologies that both benefit and protect citizens and democracy, and individual citizens have agency in protecting themselves and their communities.

- Deepfakes challenge existing legal frameworks in areas such as defamation, intellectual property, and privacy rights; and there is currently little distributor liability for social media platforms circulating deepfake content. Adapting and updating laws to account for the unique challenges posed by deepfakes while clarifying issues related to accountability, liability, and the rights of individuals affected by deepfake manipulation should be prioritized.

- Fostering research and development of technologies that can detect and mitigate deepfakes is a significant policy consideration. Increased collaboration with industry experts to establish standards and guidelines for responsible use of deepfake technology is essential, as is balancing innovation with necessary regulations to address the risks while fostering technological advancements.
- A promising approach is that of content authentication. Rather than trying to 'detect' AI generated content, the architecture of 'authentication' is instead embedded into the framework of the internet itself via a cryptographic marker embedded in the 'DNA' of the content.
- Societal norms and discourse on deepfakes should facilitate an environment where people are skeptical about what they see and are encouraged to challenge each others' informational claims. Digital literacy training, especially if directed at societal thought leaders and influencers, assists in increasing awareness of risks as well as trust in media.

That deepfake technology will continue accelerating towards producing more realistic content more efficiently and more cost-effectively is a certainty. Considering deepfakes from a global perspective allows for comprehensive approaches to maximize the benefits of the evolving technology while addressing the associated individual and national security risks, upholding privacy rights, and maintaining public trust in media and information sources.

Deepfakes: A Real Threat to a Canadian Future

The act of creating and/or sharing false, misleading, or sensationalized information is far from novel, with instances dating back to as early as the 15th century[1,2]. However, due in part to a recent (2016) resurgence in right-wing nationalism, "fake news", specifically disinformation, has been deemed an issue of concern by both academia and the public[3].

Disinformation is false information that is deliberately intended to mislead[4]; it is not only inaccurate but intends to deceive and inflict serious harm[5]. The internet and social media facilitate the speed at which present day disinformation can spread, as well as the sweeping magnitude of influence it can have. There is also another advantage that present day disinformation has over traditional forms: the existence of deepfakes.

Deepfakes are media manipulations that are based on advanced artificial intelligence (AI), where images, voices, videos or text are digitally altered or fully generated by AI[6]. This technology can be used to falsely place anyone or anything into a situation in which they did not participate — a conversation, an activity, a location, etc[7,8]. AI-generated text such as articles, blogs, and reviews, whether truthful or not, can be quickly posted online amongst 'real' content[9].

While deepfake technology is being used to create wholesome, entertaining, and satirical content, governments need to consider the potential harms and/or threats to public safety that this technology poses. As this paper demonstrates, deepfakes warrant the attention, and action of democratic governments and those who value the freedoms and safety of life in a democratic country.

Advancements in AI

The Treasury Board of Canada (TBS) defines AI as information technology that performs tasks that would ordinarily require biological brainpower to accomplish, such as making sense of spoken language, learning behaviours, or solving problems[10]. In simpler terms, AI refers to a computer performing tasks that humans do; for example, speech recognition, decision-making, identifying objects, or translating from one language to another[11,12].

In the last few years, AI has advanced significantly in its performance of these so-called "human tasks". ChatGPT (openai.com) is a prime example of such an advancement[13]. The "GPT" references generative pre-trained transformer, which is the type of large language model on which ChatGPT is built. Unlike a typical chatbots, ChatGPT communicates using "humanlike" dialogue. This means that it generates responses based on the context and tone of the "conversation" it is having with users. ChatGPT also (usually) provides accurate answers to the questions that users pose, since it has been trained on data taken from the internet[14].

Text-to-image generation is another significant advancement for synthetic media, wherein an AI model creates a unique image based on user-inputted key words[15]. With enough training data, and some additional advancements, it is likely that these models will soon be able to generate images that place real people in visually realistic, but completely fabricated, scenarios.

AI has also advanced the speed at which human tasks are accomplished generally. AI-based program AlphaFold, for example, is assessed to have correctly predicted the structures of over 200 million proteins[16], greatly advancing research on likelihood and treatment of disease. Without AI, this revelation would have taken human researchers years or decades to achieve[17].

The Threat to a Canadian Future

Advancements in AI are rapidly improving the realism of deepfakes, as well as making them more difficult to detect and disseminate[18]. Deepfake applications are also becoming more accessible and less technical[19]. Despite this, there appears to be a lack of awareness or knowledge around deepfakes, and an inability to recognize or detect them[20].

Taking these factors into account, the question of whether deepfakes are an issue of concern for Canadians is raised. In response, one could consider a statement made by Nobel Laureate Maria Ressa:

> *Without facts, you can't have truth. Without truth, you can't have trust. Without trust, we have no shared reality, no democracy, and it becomes impossible to deal with our world's existential problems*[21, 22].

Ultimately, the issue centers on facts. If a democratic society is unable to differentiate fact from fiction, then how is it going to survive? How is Canada going to function if there are different sets of unverifiable facts that different segments of the population believe in? If disinformation is unmanageable and/or unidentifiable, how is Canada going to develop solutions for those real problems? What does this mean for Canadian values and/or the Canadian way of life? Moreover, what happens when deepfakes are used for malicious purposes or with the intent to harm Canadians or their allies?

Harms Caused Using Deepfakes

The harms caused by deepfakes are illustrated by recent examples. In January 2023, a young woman named Blaire, a Twitch streamer and YouTuber better known as 'QTCinderella', discovered that a deepfake porn site was using her face/likeness in porn videos, along with the faces of other female Twitch streamers[23]. A fellow Twitch streamer, Brandon Ewing, had paid the website for deepfake porn of Blaire and other female Twitch streamers[24, 25].

In 2019, Rana Ayyub, an investigative journalist for the Washington Post, spoke out against a political party in India that was protecting the rapist of an eight-year-old Indian girl[26]. In response, a deepfake porn video of Ayyub was produced, which went viral within 48 hours. Following the release of the deepfake, Ayyub received death threats, as well as racist and misogynistic comments. Not surprisingly, for a period of time, Ayyub completely disappeared from social media and stopped reporting[27].

Such examples of deepfake porn are not uncommon. Over 90 per cent of deepfakes available online are non-consensual pornographic clips of women; as of October 2022 there were over 57 million hits for 'deepfake porn' on Google alone[28]. Women are almost always the non-consenting targets or subjects of pornographic deepfake videos, and current legislation offers victims little protection or justice[29, 30].

Notably, not all harm caused by deepfakes is of the pornographic variety. For example, criminals have repeatedly used deepfakes of Elon Musk in fraudulent cryptocurrency giveaways, resulting in financial losses totaling in the millions of dollars[31 32]. There have also been instances of scammers using voice clones of senior representatives of banks and other high net worth companies[33]. The scammers would call these offices, posing as CEOs and managers of the respective corporations, and instruct staff to initiate money transfers into their bank accounts[34, 35].

Following the COVID-19 pandemic, the world has adapted to using virtual platforms to host meetings, interviews, classes, etc[36]. With deepfake technology becoming more and more available and accessible, it is becoming increasingly difficult to verify the true identity of the individual on the screen. In 2022, the FBI noted that criminals are using deepfakes in virtual job interviews for remote jobs where the hiring company would likely only ever engage with the employee on a virtual platform[37, 38].

Other AI Considerations

It is clear that AI is a powerful tool that can advance solutions and facilitate positive outcomes to problems. However, it can also equip an entity with the power to cause significant damage. Consider the following:

1. Privacy Violations. AI systems can collect, process, analyze, and store significant volumes of data. If a system is hacked or a security breach occurs, this data (e.g., medical history, biographic data, and/or banking information) can easily be stolen, manipulated, and/or extorted for nefarious purposes[39,40,41].
2. Social Manipulation. AI can be used to track, analyze, and predict individuals' online activities, which can make them vulnerable to manipulation or being compelled to engage in activities in which they would not otherwise participate[42, 43].

3. Bias. Humans, with their implicit and explicit biases, build, train, and test AI systems. These biases can be reflected in the decisions taken by AI systems, to the detriment of certain groups[44, 45].

A Path Forward for Governments

Globally, an estimated 5 billion people use or have access to the internet[46], and Canadians make up 36 million of these users[47]. This means that AI-driven technology and disinformation is accessible to, and can potentially influence, a significant portion of the Canadian and global audience.

Deepfakes and other advanced AI technologies threaten democracy as certain actors seek to capitalize on uncertainty or perpetuate 'facts' based on synthetic and/or falsified information. This will be exacerbated further if governments are unable to 'prove' that their official content is real and factual.

Deepfakes and synthetic media can also facilitate psychological, reputational, and economic harms[48, 49]. As previously noted, the use or exploitation of AI systems can result in privacy violations, social manipulation and/or harm caused by inherent bias. Governments have a responsibility to intervene in such threats to its citizenry.

Indeed, a number of the foremost experts in AI—Yoshua Bengio, Elon Musk, and Geoffrey Hinton—have all emphasized the dangers that AI presents to the public[50, 51, 52, 53]. Hinton even resigned from his position at Google so that he could speak more freely about these dangers[54, 55]. The level of concern exhibited by these experts, coupled with the potential for harm to citizens, should all but demand that governments address AI and its potential impacts on its citizens.

Canada's TBS has indicated that the Government of Canada is committed to using AI to support and/or improve some of the services it provides to Canadians in a manner that is compatible with the core principles of administrative law. This is the basis for *TBS' Directive on Automated Decision Making,* which aims "to ensure that automated decision systems are deployed in a manner that reduces risks to clients, federal institutions and Canadian society, and leads to more efficient, accurate, consistent and interpretable decisions made pursuant to Canadian law"[56].

Another positive step forward is the Department of Canadian Heritage's Digital Citizen Initiative (DCI), "a multi-component strategy that aims to support democracy and social inclusion in Canada by building citizen resilience against online disinformation and building partnerships to support a healthy information ecosystem". The DCI aims to help both Canadians and the Government of Canada to better understand disinformation, and its impacts in order to determine the appropriate actions to take and inform future policies[57].

AI capabilities will continue to advance and evolve; the realism of deepfakes/synthetic media is going to improve; and AI-generated content is going to become more prevalent. This means that governmental policies, directives, and initiatives (both present and future) will need to advance and evolve in equal measure alongside these technologies, including capacities to characterize and differentiate malicious AI-based content from prosocial and positive applications.

When it comes to drafting and implementing new policies, procedures and/or legislation, democratic governments are, perhaps by necessity, notoriously slow moving[58, 59, 60]. AI, in stark contrast, advances and evolves rapidly. If governments assess and address AI independently and at their typical speed, their interventions will quickly be rendered irrelevant. Collaboration amongst partner governments, allies, academics, and industry experts is essential to both maintaining the integrity of globally distributed information and addressing the malicious application of evolving AI.

Disinformation, Deepfakes, and the Human Response

The realism and vividness of deepfakes makes them unusually effective at depicting alternative people and facts. People share deepfakes not necessarily because they believe them, but because they want to reinforce their identity and social position. Deepfakes rarely change minds; the threat they pose is to radicalize by sowing chaos and confusion. The following reviews the consequences of deepfakes in both the social sphere and people's private lives and suggests potential interventions to reduce these negative consequences[61].

Deepfakes are hyper-realistic digital impersonations or falsifications of images, video, and/or audio created through neural networks using machine learning models called generative adversarial networks (GAN). They appear in a wide range of contexts, from arts and entertainment to advertising, and education. The most frequent application of deepfakes is in pornography; as of October 2019, 96 per cent of deepfakes on the internet were pornographic[62]. Nevertheless, it is deepfakes' actual and potential contribution(s) to the epidemic of fake news that has engaged academia and the media, although they also pose a number of social threats[63]. Images and videos carry meaning by appearing to represent real life directly. Numerous deepfakes have appeared of politicians making claims that contradict their actual position, like the one uploaded to a hacked Ukrainian news website that depicted Volodymyr Zelensky, President of the Ukraine, telling his soldiers to lay down their arms[64] or Barack Obama using profanity towards Donald Trump[65]. If such deceptions go viral, they could have an irreversible effect on world affairs.

Humans process visual data naturally and thus fluently[66], and people believe what they see[67]. Moreover, the detailed imagery of deepfakes has the potential to prime psychological proximity. Concrete misinformation (including disinformation) primes participants to think of events as being nearer and more probable, increasing their perceived threat[68], and likelihood of news about them being shared[69].

How Effective Are Deepfakes?

Studies disagree about whether people are able to discriminate deepfakes from real images. When compared to traditional sources of fake news, such as text or audio, some studies have found deepfakes to be no more credible or effective at implanting false memories than their counterparts[70, 71]. However, while these studies are recent, the technology is changing so quickly that the studies did not deploy the most recent AI-based image-creation technologies available online. Worse, some studies of deepfake credibility use only a single video[72].

Lago et al. (2022) reports that newer AI-synthesized images are perceived as real[73]. Indeed, synthetic faces generated by the most state-of-the-art GANs were judged as more real than actual real images, pointing to the potential of deepfakes to simulate reality and circumvent the eerie, unsettling feeling that arises when humanoid robots or computer-generated images are too close to the real thing (the "uncanny valley" effect)[74]. Further, Köbis et al. (2021) show that people cannot reliably detect deepfakes, and that neither raising awareness nor introducing financial incentives improves their detection accuracy[75]. An even more recent study found that deepfake videos are both more believable than fabricated images and text and that people are more likely to engage with them[76]. Rapidly evolving GAN technology will soon render deepfakes indistinguishable from genuine content if it has not already done so.

Do People Care About Accuracy?

Information veracity is not a deciding factor when users choose to share content online[77]. Even when users can identify deepfakes as untruthful, they still might share them within their social circle. Indeed, Vosoughi et al. (2018) found that fake news is diffused faster and further online than factual information[78].

A hint about why people share and view deepfakes can be found where they are most common. The term "deepfake" was coined by a Reddit forum created for sharing pornographic videos of women whose faces were synthetically swapped for those of others, mostly celebrities[79]. Consumers of pornographic deepfakes are unlikely to be fooled by the imagery they are watching as the website or video title is typically marked as "fake"; there is no pretense of truth. Thus, viewers of pornographic deepfakes obtain whatever benefits they get despite or because of their knowledge that what they are watching is fake. This could also be true of many political deepfakes. Other uses of deepfakes similarly do not depend on their veracity, such as their use(s) for artistic and educational purposes.

Most deepfakes on social media will exist for the same reason as fake news, to attract 'clicks'. To achieve this, fake news exploits two properties that engage people's attention: novelty and negativity[80]. Sharing novel facts holds social value because it suggests that the sharer holds inside information[81]. The tendency to attend to negative information is well established. People attend more to potential losses than to gains[82]. Sharing negative information has the veneer of nobility by warning others of potential threats. Health professionals have been found to be more willing to retransmit false rumours to prevent negative repercussions (e.g., causing cancer) than to produce positive outcomes (e.g., curing cancer)[83, 84].

People share with their ideological community because it satisfies a fundamental human motivation to strengthen one's social attachments[85]. It also confirms one's identity as part of an ideological group[86]. Protecting self-identity takes priority over judging accuracy[87]. Indeed, identity (e.g., nationality, religion, race, and/or political party) will colour what people consider to be true[88]. However, because people tend to live in information bubbles, partisan belief differences generally demonstrate an ignorance of inconvenient truths rather than an acceptance of falsehoods. Therefore, people may see and share a deepfake video that aligns with their beliefs while never coming across any reason to believe that it is in fact a deepfake. In one study showing subjects faked photographs, conservatives were more likely to "remember" Barack Obama shaking hands with the president of Iran, while liberals were more likely to "remember" George W. Bush on vacation with a celebrity during Hurricane Katrina (neither event actually happened)[89]. Deepfakes have been shown to radicalize people against the opposition[90], therefore, like other information sources, deepfakes may be more likely to radicalize existing views than to change people's opinions.

Older people are more likely to be deceived by deepfakes, and political ideology influences how deepfaked news is evaluated[91]. While Republicans and Democrats in the US are equally inclined to share fake news[92], low-conscientiousness conservatives (e.g., those least likely to follow societal norms for impulse control) are the most likely to share misinformation due to their desire for chaos[93].

Consequences

Societies, companies, and consumers are all potentially threatened by deepfakes. Caldwell et al. (2020) ranks fake audio or video content as the single biggest threat posed by AI for applications to crime and terrorism[94]. Europol (2022) has warned that deepfakes can be used to harass and humiliate people online, perpetrate extortion and fraud, falsify online identities and fool “know your customer” mechanisms, sexually exploit children online, falsify or manipulate electronic evidence for criminal justice investigations, and disrupt financial markets[95].

Deepfakes also pose a threat to our governing structures. The uncertainty deepfakes introduce allows people to live in their own subjective realities, enlarging social divisions and obstructing the democratic process[96]. This is especially dangerous during elections when deepfakes are likely to be used by both foreign and domestic powers to manipulate outcomes[97]. Antagonistic parties may be enticed to subject an electorate to deepfakes long before an election in order to prime future attitudes[98].

In regards to businesses, threats include fake reviews of consumer items, defamation and sabotage, and damage to a firm’s image, reputation, and trustworthiness[99]. For instance, in 2019 criminals successfully impersonated the head of a firm’s parent company with voice spoofing software thereby tricking the CEO of a UK energy company into transferring $243,000 USD to them.

Beyond such “deepfake phishing”, AI will render some technologies obsolete. Threats to consumers include new susceptibility to blackmail, intimidation, sabotage, harassment, defamation, revenge porn, identity theft, and bullying.

The personalized nature of deepfake pornography adds a new layer of emotional distress and threat for victims[100]. Most pornographic deepfakes present celebrities whose reputations may provide a degree of shelter from being seen as the genuine subjects of the videos. They also possess public platforms, as well as legal and financial means to dispute the veracity of the videos. In cases of revenge porn, private citizens do not have even these limited protections[101].

Even when citizens do not believe the misinformation presented to them or are not concerned about truth, deepfakes can increase uncertainty about content and decrease trust in media[102]. In the US, fake news caused 50 per cent of Republicans and 38 per cent of Democrats to reduce the amount of news they consume[103]. As COVID-19 exemplified, in times of crisis this atmosphere of conspiracy and uncertainty can leave citizens vulnerable to misinformation[104]. Deepfakes are exacerbating this problem.

The use of deepfakes against public individuals creates the Liar's Dividend: individuals facing accusations can write off factual evidence as deepfakes[105]. Widespread deepfakes can prime individuals into doubting the authenticity of information. The Malaysian Minister of Economic Affairs deflected evidence of his involvement in a sex tryst by proclaiming it as a deepfake despite no evidence[106]. More recently, Elon Musk's lawyers used it in a lawsuit[107].

Deepfakes offer plenty of potential benefits. They have enabled new and intriguing art forms, served as excellent pedagogical tools, and been a benign source of pleasure and amusement. They can also offer business opportunities[108]. Facebook's metaverse will be largely composed of deepfake objects. Deepfakes afford new forms of marketing campaigns (e.g., through the removal of language barriers), virtual brand ambassadors (Lil Miquela is a fake influencer who has over 3 million followers), and a range of technical innovations. To illustrate, there are now virtual newsreaders based on real people. Deepfakes can also be used to enhance memory by, for example, making a dead person seem alive.

In the wrong hands, deepfakes are a novel kind of social virus, and like all viruses, their future trajectory and consequences are hard to predict. On a societal level, their greatest threat is their ability to shape public discourse. When misinformation enters the public conversation, it becomes increasingly dangerous as it alters collective understanding and memory. Their increasing prevalence could also lead people to stop believing much of what they see.

Solutions

Deepfakes are created to trick us; the human mind is not prepared to always accurately identify the outputs of sophisticated technologies. While some tech giants have started flagging some content as misinformation, such flags are not a silver bullet. The shareability of fake news has been found to decrease when it is accompanied by warnings[109]; however, their effect on its believability is unclear[110]. Prior exposure to misinformation increases its perceived accuracy, possibly negating the effectiveness of tags. Detection systems integrating both human and model predictions have been found to be more accurate than humans and automatic detection methods working alone[111].

Steps that might alleviate the problem include pre-exposure warnings that make people aware that information might be false before they see it. Warnings need to be specific; it is ineffective to merely mention that misinformation may be present[112]. In addition, warnings should come with an alternative causal account that explains both what happened and the reason for the misinformation. Companies can educate consumers about their products, brands, and services, helping them identify firm-sponsored and credible sources of information[113].

From a legal standpoint, there is currently little distributor liability for social media platforms circulating deepfakes[114]. In the United States, the legal debate is centered around Section 230 of the Communications Decency Act, which prevents companies from being held liable for the content on their platforms[115]. The justice system could specify civil liability for the creators and distributors of deepfakes, while also increasing legal protection for victims of defamation.

Individuals have little power to prevent deepfake attacks. When deepfakes threaten reputations, individuals can increase their ability to deny actions by recording their activities, but this raises privacy concerns[116]. Methods to disseminate facts can help protect communities if they are deeply informed by understanding of the public's information landscape and means of navigating it.

Individuals are more easily persuaded and corrected by someone they know. Therefore, societal norms and discourse on deepfakes should be nudged to create a social environment where people are not only more skeptical about what they see, but also are encouraged to challenge each others' informational claims.

To alter societal norms, thought leaders and those most central in social networks are key. Educational resources including digital literacy training are helpful tools, especially if directed at influencers. Videos explaining political deepfakes have been found to reduce uncertainty, and in so doing can increase trust in media[117]. But norms only really change through collective action.

Real People Using Fake People: Public Use of Deepfake Technology

Synthesizing realistic audio, images, and videos using algorithms has always been essential in Signal Processing, Computer Graphics, and Computer Vision. When using pre-artificial intelligence (AI) tools, this process is usually lengthy, costly, and technically demanding for ordinary users. However, the rapid developments in AI technology in recent years have significantly lowered the resources, time, and technical expertise required to create compelling fakes. Such developments first caught the public's attention in late 2017 when a Reddit account called 'DeepFake', a portmanteau of deep learning and fake media, began spreading pornographic videos with transplanted celebrity faces created using a Deep Neural Network (DNN)-based algorithm. Since then, more sophisticated algorithms that synthesize realistic audio, images, and videos have emerged, along with a plethora of open-source software tools and commercial services. 'Deepfake' is also used more broadly as a term that refers to any impersonating media created or edited by deep learning algorithms.

Deepfakes are just the tip of the iceberg of this troubling trend. By creating illusions of an individual's presence and activities that did not occur in reality, deepfakes can cause real harm when they are weaponized. For instance, a fake video showing a politician engaged in an inappropriate activity may be enough to sway an election if released close to voting day. A falsified audio recording of a high-level executive commenting on her company's financial situation could send the company's stock into freefall. Using a synthesized realistic human face as the profile photo for a fake social platform account can significantly increase the impact of deception. An online predator can masquerade as a family member or friend in a video chat in order to lure unaware victims. Left unchecked, deepfakes can escalate the scale and danger of online disinformation and fundamentally erode society's trust in digital media.

Recent developments in generative AI are significantly boosted by the availability of large-scale language and image generation models, such as the OpenAI Generative Pre-trained Transformer (GPT) family, DALL-E, and Midjourney. Catching the public's imagination of the superpower of AI technology and hinting at the prospect of Artificial General Intelligence, these developments have also opened up new opportunities and challenges in the making of deepfakes. These developments have focused on three main directions: i) making the models more powerful, and capable; ii) making them more accessible; and iii) giving users more control over the style and content of the generated media through detailed text prompts.

One of the most significant advancements in generative AI is the increased power and capability of the models. This has been made possible by the availability of vast volumes of training data that enable the models to learn complex patterns and generate high-quality output. These models can generate realistic and complex images, videos, and audio that are almost indistinguishable from those created by humans. The applications of these models are vast, ranging from generating realistic images for virtual environments to creating realistic voices for virtual assistants.

Another important direction in generative AI is the accessibility of the models. Many tools now provide web-based interfaces that require little to no coding and/or installation effort, making it easier for non-experts to use and benefit from these models.

Finally, developing AI tools that give users more control over the style and content of the generated media through detailed text prompts is another important direction. This enables users to specify the desired output style and content by providing text prompts that the AI model can use as input. This can be useful in generating customized content for marketing campaigns, creating personalized content for social media, or generating realistic simulations for training purposes.

The main forms of current deepfake making methods are summarized in three categories: images, video, and audio/voice.

Images

A quintessential example of deepfakes is the highly realistic images created from the generative adversarial network (GAN) models. A GAN model consists of two DNNs trained in tandem. The 'generator' synthesizes images, and the 'discriminator' differentiates synthesized images from real ones. In training, the two DNNs compete: the generator tries to create more realistic images to defeat the discriminator, while the discriminator attempts to improve the classification accuracy. The training ends when the two DNNs reach an equilibrium. The generator is then used to create realistic images from input white noises.

Recent works, known as StyleGANs, have demonstrated the superior capacity of GAN models in generating high-resolution and realistic human faces. GAN models can also be used to edit or transfer the attributes and expressions of faces. A more recent variant of the image generation model is known as the diffusion model. Like the GAN model, the diffusion model creates realistic images from input noise. However, the training mechanism of the diffusion model is different. It uses a Deep Neural Network to simulate the physical process of diffusion, in which a structured signal is slowly dissolved into thermal-dynamical equilibrium through the stochastic process of diffusion—imagine a drop of ink dissolving in a cup of water. The deep neural network is then used as the reverse model to transform input noise into a structured image. Diffusion models have led to the state-of-the-art generation of realistic human faces with software systems such as Stable Diffusion being widely used.

Videos

The original namesake of 'DeepFake' is face-swap videos generated using an image-to-image translation framework. Specifically, the faces of a target are replaced by the faces of a donor synthesized using the auto-encoder (AE) model. The AE model consists of two DNNs, encoder and decoder, trained using the target and the donor's faces. The encoder retains the target's facial expressions and head poses while the decoder combines these with the target's identity. This approach of synthesizing face-swap videos has been mainstreamed through open-source software implementations on GitHub (github.com).

There are also techniques to create videos of upper-body reenactment, and whole-body motions. Other variants of this method are those which animate a single face image from a driving video of another person. Examples of such methods are Reenact GAN and First Order Motion. These methods use Deep Neural Network models to transfer the facial movement from the driving video to the input face image to create a video sequence of the subject in the image with the same facial movement. Several start-ups have commercialized the making of face-swaps or reenact videos (Synthesia and Canny AI, for example).

Audio/Voices

DNN models have also been used to create realistic, synthetic human voices. Two types of deepfake audio differ in their input modality. The text-to-speech models (e.g., Parrotron and Spectron) convert an input text to the target's voice, while the voice conversion models use a source person's voice as input. The underlying speaker-adaptive neural speech synthesis system usually includes i) acoustic modeling models, ranging from simple spectrograms to the more sophisticated neural speaker and style embedding (e.g., Tacotron and its variations); ii) vocoders such as WaveNet or WaveRNN for speech waveform generation; and iii) conversion algorithms based on auto-encoder or GAN models. Several commercial companies, such as Lyrebird, Respeecher, Murf.ai, ElevenLabs, and Dessa, provide voice imitation as a service.

Multimodal Generation

Text-to-image generation has significantly improved in the past two years with recent advancements in attention-based transformer and diffusion models. Several large-scale language-image models have been developed, including the DALL-E model proposed by OpenAI in 2021, which uses an autoregressive transformer to generate high-quality images on the MS-COCO dataset without any training labels. Other models such as CogView, Parti, Make-A-Scene, and most recently, MidJourney, have also used autoregressive transformer models for text-to-image generation. In 2022, an updated version of DALL-E, DALL-E2, was developed using a diffusion model with CLIP image embeddings, enabling it to produce higher quality and more diverse samples more efficiently. Other models, such as GLIDE, Stable-Diffusion, and Imagen, have also used diffusion models to improve text-to-image synthesis.

These powerful text-to-image synthesis models have inspired several studies focused on developing text-guided image editing models, including DiffEdit, Prompt-to-prompt, Null-text Inversion, Imagic, and Muse. These models perform local semantic editing to an image based on text input with the desired edit and an optional scene layout (segmentation map). However, their optimization often maximizes similarity to the original image while maintaining the ability to perform meaningful editing on local regions. This type of entire synthesis can be easily identified if seen in the training data.

AI methods that create lip-synch videos with input voices for any video have become increasingly popular in recent years. These methods aim to generate realistic mouth movements that synchronize with the audio of a person speaking in a given video, allowing the video to be dubbed or re-voiced in a different language. One common approach to lip-synch video generation is deep learning-based models that can learn the relationship between audio and mouth movements. These models usually involve training on large datasets of audio-visual pairs to learn the mapping between the audio and visual domains. Other approaches involve using facial landmark detection techniques to predict the movements of the lips based on the audio input. Recent advancements include using neural machine translation techniques to enable lip-synch generation in different languages and integrating natural language processing techniques for more accurate and contextually relevant lip-synch generation.

Conclusion

Although the future of deepfakes is hard to predict, one thing is certain, the technology will continue accelerating towards producing more realistic content more efficiently and more cost-effectively. Various stakeholders will need to take action to control the potential misuse of such tools for disinformation. The most direct measure is for the service/tool providers to regulate the uses and watermark the generated contents so that they can be traced and exposed more easily when spread on social media. Platform companies are also responsible for filtering and limiting the viral spread of synthetic content, and the associated orchestrated disinformation campaigns.

Public media can help users to expose disinformation through fast response times in fact checking and debunking. Users also need to increase their own awareness and knowledge of synthetic media and be encouraged not to spread unreliable information. Lastly, government agencies can play a critical role in guiding national research strategies to invest more into researching countermeasures to deepfakes while focusing legislative efforts to control the problem.

Commercializing AI: Applications in Tech, Industry, and Business

Recent progress in artificial intelligence (AI) does not seem to be slowing down, showcasing new technological capabilities at a rate that makes it difficult to imagine what may next significantly impact daily lives and businesses or what may become obsolete. Tech companies such as OpenAI, Google, Microsoft, Meta, NVIDIA, Apple, and Adobe are investing heavily in research and development (R&D) and adopting emerging AI capabilities into their products as quickly as possible to avoid missing any major disruptive opportunities.

Current-day discussion on AI mostly revolves around generative AI technologies and transformer-based products such as ChatGPT or GPT-4. These technologies are being incorporated into search engines like Bing (bing.com), while huge amounts of AI-generated images from diffusion models are circulating on social media. Generative AI for media synthesis can refer to the generation and manipulation of images, videos, audio, and even 3D content. Media can be generated with (text, image, audio, etc.) or without input (random noise).

The commercialization activities and opportunities of AI are limitless and span almost every industry sector (telecommunication, health, transportation, education, energy, entertainment, etc.). This paper focuses on:

- the commercialization of major generative AI technologies for media synthesis in recent history (what is real and what is not); and
- the new technological capabilities and potential disruptive products on the horizon.

Deepfakes: Still Harmless Synthetic Media

Since the emergence of so-called deepfake technologies, public concerns have been raised about potential dangers in the context of disinformation, fraud, and harassment, especially due to the technology's wide adoption for non-consensual internet pornography. Deepfake technologies are a form of media synthesis, often based on generative AI that is focused on manipulating specific human subjects in a video (e.g., through face swapping, facial puppeteering, or lip-synching). The risk of deepfake technologies being used for malicious purposes seems particularly tangible as they can generate extremely convincing results (traditionally only achievable by professional visual effects studios), can make people appear to say and do anything one wishes them to, and the technology is largely accessible to anyone as it is relatively easy to learn and requires no technical expertise.

While technologies for media manipulation have continued to progress at a steady rate, deepfakes themselves seem to be rather harmless and not as catastrophic as many experts predicted. Deepfakes have been used for online fraud and harassment cases, during political elections, or by Russian activists in the Ukraine War, but so far have been ineffective as a tool for disinformation (especially compared to fake news in general). Many of the deepfakes (images or videos) circulating on the internet are generated by non-experts, and despite looking impressive, they still appear artificial and do not require sophisticated detection algorithms to be uncovered.

Productization of Generative Media

Mobile Apps, Augmented Reality Filters, and Social Media Videos

Beyond some gimmicky and entertaining mobile apps and augmented reality (AR) filters (Snap, TikTok, etc.), deepfake technologies may not initially seem to have any deeper meaningful purpose beyond manipulating videos and adding new visual effects to social media posts. Some of the most common effects include inserting a user's face into a video clip (e.g., Zao, ReFace) by uploading a single photo and selecting a pre-curated video, and swapping the face of a user video with that of a celebrity (e.g., Impressions.ai) by uploading a video and selecting a pre-trained face model of a celebrity. Facial reenactment using a technique called first order motion has also gained popularity, where an arbitrary portrait of a person is uploaded and immediately reenacted, with users being able to create viral videos of politicians singing or reenacting people from the past (e.g., DeepNostalgia, MyHeritage).

The demand for new and more impressive tools for self-expression are driving researchers in both academia and industry to continue to push the limits of media synthesis capabilities (e.g., higher resolution, real-time performances, more control, less artifacts, more accessibility, etc). This has resulted in new, sophisticated filters such as de-aging, gender swaps, cartoon filters, and the introduction of new algorithms with innovative input modalities (single image reenactment, text-based specifications, photo-real avatars from videos, etc.).

Virtual Assistants, Marketing Videos, and Universal Translators

While essential in many entertainment applications, other commercial sectors have also explored the use of digital humans to improve, automate, and scale their services through the use of generative AI. Several companies have developed human-like virtual assistant solutions (e.g., Soul Machines, Uneeq) but they fail to appeal to customers due to their "uncanny valley" appearances (the feeling of discomfort caused by viewing imperfect computer-generated faces)[118,119]. Despite technological advances in the application of graphics engines (e.g., Epic Games / MetaHumans, unrealengine.com) or generative AI to enhance photorealism in those avatars (e.g., Samsung Neon, Pinscreen, etc.), virtual assistants still struggle to replace real humans. They currently lack sophisticated responses, and their voices and facial expressions are often absent of emotion and empathy.

However, due to recent advancements in large language models (LLM) such as ChatGPT and emerging research in motion synthesis, the mass adoption of highly convincing and realistic human-like AI agents may be closer at hand (within two to three years), especially if they have the ability to interact in real-time. In the meantime, several startup companies (e.g., Synthesia, Colossyan, etc.) are exploring the use of generated videos of pre-recorded humans in a non-interactive setting where marketing and training videos are generated at scale for enterprise applications. An actor and/or voice can be selected through a web interface, and a text script provided as input in order to generate video content automatically on a server. These solutions typically use a text-to-speech solution (e.g., third party or proprietary where voices can be customized), and a speech-to-face video generator that uses an audio input and video frames as training data (e.g., for Synthesia: ten minutes of an actor performing a speech in a well-lit studio environment and facing the camera head-on).

These methods are more advanced than the popular wav2lip algorithm and generate higher resolution and better quality results. Similar technologies have also been adopted by Chinese Tencent and Korean companies such as DeepBrain in the context of generating news anchors and marketing material at scale. Tencent for instance charges only $145 USD for each subject (either half or full body) and supports both English and Chinese languages. Despite their high level of fidelity, the resulting human performance generated from speech still appears slightly robotic during conversations, and mass adoption is still limited.

Google recently announced at their I/O Conference an enterprise-level service called Universal Translator, which allows educational content creators to translate their videos into multiple languages. The solution uses translated voices as input to generate synchronized lip movements in the videos. The translated voice input is also produced using a generative translation model that mimics the voice and tone of the speaker but in a different language. As of now, this offering is only available for select and authorized content creators (e.g., partnership with Arizona State University), which can assist in preventing its use for malicious applications.

Cheaper and Faster: Visual Effects (VFX) and Visual Dubbing for Hollywood

Whether it is to create digital stunt doubles, bring deceased stars back to life, or de-age an older actor, computer-generated digital actors are widely deployed in some of the most memorable blockbuster films (e.g., *Star Wars, Furious 7, Terminator: Dark Fate*, and *The Curious Case of Benjamin Button*). However, these effects typically rely on sophisticated visual effects studios (e.g., Industrial Light & Magic, Weta Digital, MPC, Framestore, etc.), cost millions of dollars, and take months of work to produce a few seconds of footage. Visual effects related to human facial performances are particularly expensive and difficult to achieve due to the "uncanny valley" effect.

As open source deepfake solutions (such as faceswap-GAN and Deep Face Lab) emerged and became freely accessible on the Internet, hobbyists and deepfake artists began generating entertaining videos by swapping celebrities in short video clips. While it was possible to produce highly convincing deepfakes, the resolution was often still too poor for film production. However, these methods quickly caught the attention of VFX producers as a tool for enhancing their conventional VFX pipelines in order to save cost and impact story telling. Visual effects companies such as Industrial Light & Magic (ILM) have explored the use of deepfake technologies for de-aging actors (e.g., Mark Hamill in *Star Wars*, Harrison Ford in *Indiana Jones 5*). This is achieved by replacing the faces of aged actors or doubles with neural renders built from younger footage of the same actor and combining them with 3D models and video compositing techniques.

AI startup companies such as Pinscreen and Metaphysic provide complete AI visual effects solutions for face replacement in film production. Metaphysic is known for its viral Tom Cruise deepfakes circulating on TikTok and their recent Elvis face replacement on America's Got Talent.

Pinscreen innovated the development of a number of GAN-based neural face rendering technologies (most notably PaGAN, "photoreal avatar GAN"), which were originally developed to enhance the realism of 3D avatars for interactive 3D and metaverse applications. In 2022, the company started to shift its focus in the VFX space through a partnership with Netflix and Amazon Studios, and launched on a number of high profile TV shows (e.g., *The Manifest*), blockbuster movies (e.g., *Slumberland 2022*), and advertisements (Nike, Balenciaga, etc.) using generative AI technologies. AI VFX services include end-to-end processing for face replacement, facial reenactment, aging/de-aging, and visual dubbing. Pinscreen's key advantage consists of being able to handle very short cinematic shots and deliver high-fidelity 4K HDR output, allowing for the processing of close-up shots, extreme side views, and dramatic/dynamic lighting conditions.

The process requires specialized GAN-based data augmentation and AI enhancement procedures to generate unseen data from sparse views collected from film footage and improved architectures for high resolution and temporally coherent video synthesis.

Despite the growing demand of AI VFX services such as face replacement, aging, and de-aging, these remain relatively niche applications and are highly show-dependent. One scalable market is in visual dubbing for films and TV shows, where foreign films can be watched in any desired language while also having actors' lip movements perfectly synchronized to speech. Feature films are much harder to process than video clips that are captured in controlled settings (such as of news anchors, marketing and training materials) due to the complexity of scenes, lack of training data, and the extremely high quality requirements in cinema (4K HDR).

In 2022, Pinscreen became the world's first company to fully lip sync a full feature film using its proprietary generative AI pipeline, demonstrated on the film The Champion — translated from German/Polish to English. The process combined state-of-the-art generative AI and an integrated VFX pipeline, which allowed Pinscreen to complete the processing of a 90 minute film in less than three months. The approach can handle an existing movie, and only requires additional video recordings of the voice actors during the dubbing process. Other players in AI VFX such as Flawless.ai are trying to enter the market for visual dubbing but have limited technical capabilities as they only offer speech-to-face reenactment as opposed to video performance as input. They have demonstrated some visual dubbing examples on select video clips, but not entire movies.

Diffusion-Based Text-To-Image Generation

With breakthroughs such as OpenAI's Dall-E and recent advancements in diffusion and transformer-based models such as Stable Diffusion, image generation capabilities that outperform traditional GAN-based methods in terms of image quality, resolution, and diversity are now possible. The latter property is particularly significant as it enables highly effective text-to-image generation, where users can input an arbitrary text prompt allowing the model to generate an image that reflects this prompt accurately. Incorporating text input is typically enabled by using a CLIP-encoder that can map the prompt into a text embedding which is then used as a condition for generating an image using a progressive de-noising process (the generator), which is typically based on iteratively using a deep neural network based on a U-net architecture for image-to-image translation.

While training those models is easier and more reliable than GANs, diffusion model training is extremely resource intensive, typically requiring weeks of training and hundreds of high performance GPUs (A100s). Consequentially, those models are often trained by companies who have large GPU resources (e.g., OpenAI, Stability.ai, Google, etc.), while labs in academia and smaller companies rely on pre-trained models they can further fine-tune. The latest and most popular commercial solutions include OpenAI's Dall-E-2, Midjourney (via Bot on Discord), as well as Stability.ai's solution (available as web interface Dream Studio) and APIs. While incredibly realistic images can be generated, they are still prone to noticeable artifacts, and production-level fine control is not yet possible. Some level of control through scribbles or abstract skeletons have been recently demonstrated (e.g., ControlNet), but the generated images always come with unpredictable details and appearances. As a result, diffusion-based methods are not yet suitable for production-quality video generation as they lack controllability and temporal consistency.

Summary and Future Capabilities

Generative AI capabilities for media synthesis (images, video, audio) are constantly evolving. Generated image qualities are improving (e.g., higher resolution, less artifacts, and more semantically realistic results) and are more diverse, enabling natural text prompts as input. Similar to when GANs were introduced, the research community is focusing on enabling better controllability, more predictable outputs, and temporally consistent generations for videos, as well as the ability to handle other modalities such as neural 3D content. Due to this technology's accessibility and performance in generating convincing content, concerns around its potential misuse have been raised by the public. So far, these media synthesis technologies and deepfakes have not been extensively weaponized, even though they pose a potential threat.

In the coming years, society is expected to witness further technological breakthroughs in generative AI. These breakthroughs will enable new commercial opportunities, including general online video generation services (e.g., a YouTube that can take any text prompt as input and generate the desired video on-the-fly), real-time and fully interactive videos (e.g., advertisements that can interact with a viewer in real-time), as well as fully immersive and photorealistic AI-generated environments for Metaverse applications. With the recent announcements of new augmented reality/virtual reality (AR/VR) headsets such as Apple's Vision Pro and Meta's MetaQuest 3, it is foreseeable that the demand for sophisticated 3D content will grow and generative AI will play a key role in enabling content creation.

Implications of Deepfake Technologies on National Security

Society is on the verge of an era where reality can be manipulated, truth can be distorted, and trust can be shattered. Deepfake technologies pose a grave and imminent threat to national security. There are a variety of deepfake technologies available today, with differing degrees of complexity and ease of use. Some commonly used deepfake tools include Face Swapping[120], Lip Syncing[121], Voice Cloning[122], and GAN-based deepfakes[123].

Adversaries, whether state actors, criminals, or clandestine organizations, are continuously working to refine their skills in the art of public manipulation. Deepfakes can further enable these adversaries to achieve their goals of exploiting vulnerabilities, sowing discord amongst the public, and/or undermining the very essence of democracy.

This paper provides an overview of the capacity, impact, and underestimated dangers of deepfakes while recognizing the urgent need for robust and holistic mitigation strategies in the face of this rapidly evolving and highly sophisticated threat.

Implications, Scenarios, and Consequences[124]

The capacity for deepfakes to demolish reputations in a single instant exemplifies their insidious nature. Strategic dissemination of a single fabricated video can cause public indignation, financial losses, and irreparable harm to individuals and organizations. Consider a deepfake video in which a prominent corporate CEO makes racist comments. The repercussions of such a planned assault could result in a catastrophic drop in the company's stock price, eroding not only its financial stability but also its credibility in the eyes of investors and stakeholders. Deepfake pornography has already caused harm; a 2019 study found that 96 per cent of the 14,000 deepfake videos found online were pornographic, a number which is infinitesimally small compared to the current landscape featuring non-consensual pornography depicting high-profile individuals such as journalists and celebrities[125].

The malicious potential of deepfakes extends beyond external threats; so-called 'insiders' can provide the ultimate advantage to adversaries and/or bad actors. An employee with access to sensitive information can utilize deepfakes to facilitate the leaking of classified data or engage in other illicit activities, thereby compromising national security, jeopardizing the organization's integrity, putting other employees at risk, and/or causing substantial financial losses. Robust employee screening and monitoring protocols must therefore be implemented to identify and address potential insider threats. Moreover, the development of advanced deepfake detection technology tailored specifically for insider threat prevention is imperative. Those who exploit their positions of trust must face severe penalties to deter others from doing so.

Deepfakes can be incorporated into social engineering campaigns in order to manipulate individuals and organizations for malicious purposes. A deepfake video claiming to show a loved one in peril and/or requesting a large sum of money, for instance, can result in significant financial loss for the target. Increasing awareness of social engineering that deepfakes can facilitate, developing effective detection and response protocols, and nurturing cooperation between law enforcement agencies are essential mitigation strategies.

The economic implications of deepfakes extend beyond reputational damage, as they have also emerged as an effective tool for economic espionage. Adversarial entities can exploit deepfakes to target businesses or industries, aiming to gain a competitive advantage or disrupt critical sectors of the economy. Mitigation strategies include implementing effective deepfake detection technology for business communications, enhancing media literacy for business leaders, and establishing clear protocols for responding to deepfake-related economic espionage. Further, deepfakes can be used to facilitate financial fraud by impersonating individuals with access to sensitive financial information. Implementing multi-factor authentication for sensitive financial transactions is an effective mitigation strategy.

Deepfakes have the potential to penetrate a nation's most critical, and secure systems, which poses a grave threat to cybersecurity. For example, criminals can exploit deepfake technology to bypass biometric authentication systems, gaining unauthorized access to secure facilities and sensitive personal information. The consequences of such breaches are far-reaching, jeopardizing not only individuals' privacy and also threatening national security. To counter this, anti-spoofing measures for biometric authentication systems must be implemented, ensuring that digital fortresses remain impenetrable.

Deepfakes have also emerged as clandestine weapons, allowing covert operations to be carried out undetected. False evidence can be fabricated or surveillance footage manipulated, undermining confidence in visual records, and impeding intelligence operations. To preserve the integrity of intelligence operations and bolster national security, forensic techniques that can detect deepfakes in video evidence are indispensable.

National Security and Intelligence Implications

Increasing awareness of deepfake threats and mitigation strategies among individuals and organizations at high-risk of extortion or coercion, as well as developing effective cybersecurity protocols, are crucial for preventing and/or mitigating deepfake-related breaches.

As deepfakes are capable of fabricating fraudulent evidence or manipulating public perception, they can lead to a distortion of the truth. Increasing transparency and accountability in the use of deepfake-related evidence, developing detection and verification systems, and promoting fact-checking and verification procedures in media reporting are essential for addressing this challenge.

Disruptions of critical infrastructure or government agencies, can have severe consequences. As such, implementing effective cybersecurity protocols, raising public awareness of deepfake-related threats, and conducting regular deepfake detection and response training exercises for government agencies are key mitigation strategies.

Deepfakes can be utilized as part of cyber warfare campaigns to target critical infrastructure, disrupt government operations, and/or create havoc in financial markets. Increasing awareness of deepfake-related cyber warfare threats among government agencies and critical infrastructure providers, developing effective detection and response protocols, and fostering international cooperation are crucial for preventing deepfake-related cyber attacks. Deepfakes can likewise damage diplomatic relations and/or delay treaty negotiations. Mitigation strategies include developing effective deepfake detection technology for diplomatic communications, increasing media literacy for diplomats and government officials, and establishing clear communication channels for responding to deepfake-related diplomatic incidents.

Terrorist organizations surely recognize the potential of employing deepfakes in the spread of propaganda and coordination of attacks. Even in the absence of deepfakes, terrorism jeopardizes the safety of innocent lives and the stability of critical infrastructure. Therefore, heightened awareness of deepfake-related terrorist threats among law enforcement and intelligence agencies, the development of effective detection and response protocols, and a resolute dedication to international cooperation are vital for countering this nefarious use of deepfakes.

Deepfakes can also spread disinformation specific to military capabilities or movements, potentially resulting in military conflicts. Raising awareness of deepfake-related military disinformation threats among military and intelligence agencies, developing effective detection and response protocols, and promoting international cooperation are essential for preventing deepfake-related military conflicts.

Implications for Democracy

The use of deepfakes to spread false information or propaganda can result in confusion and distrust amongst the public. Developing advanced deepfake detection technology, increasing media literacy education, and enforcing penalties for the spread of malicious deepfakes are crucial in combating misinformation.

Deepfakes can interfere with the democratic process through the manipulation of public opinion, which can impact and/or sway election results. To raise awareness of deepfakes and online disinformation and encourage political parties to work together to prevent deepfakes from impacting the 2019 UK election, advocacy group Future Advocacy created and disseminated a video in which candidates Boris Johnson and Jeremy Corbyn endorsed each other[126]. Developing effective deepfake detection technology for political campaigns, increasing media literacy education for voters, and enforcing penalties for the spread of malicious deepfakes during elections are important mitigation strategies.

Deepfakes can easily spread disinformation targeted against specific government officials or departments, which can lead to public distrust and legal action. Implementing effective deepfake detection technology for government communications, increasing media literacy for government officials, and establishing clear protocols for responding to deepfake-related disinformation campaigns are essential.

Deepfakes can be used to create convincing propaganda videos for the purpose of influencing and/or swaying public opinion. Increasing transparency, and funding for election commissions to investigate, and prevent the spread of deepfakes is critical. Similarly, in the context of influence operations, developing robust fact-checking processes, educating the public about the risks of deepfakes, and promoting critical thinking are crucial.

In addition, deepfakes can manipulate (or create) evidence in criminal investigations or legal proceedings, which can lead to wrongful convictions. Therefore, there is a need to develop guidelines for the use of digital evidence and establish procedures for the verification of video evidence in legal proceedings.

Containment and Mitigation Strategies

Containment and mitigation strategies can be technological, legal, and/or societal in nature, and depend on collaboration and responsible governance.

Deepfake detection technologies need to be developed and enhanced. Machine learning algorithms can be used to detect inconsistencies in videos, audio, or images. Blockchain technology could also be used to create immutable records of media content, making them difficult to tamper with.

Legal approaches that will criminalize the creation and dissemination of deepfakes for malicious purposes (beyond the legal repercussion of “fraud”) require consideration. Such laws should provide protections for individuals whose reputations have been damaged by deepfakes.

Media literacy and critical thinking need to be fostered and promoted amongst the public. Public education on the identification of deepfakes and the associated potential risks and impacts would be of value. Journalists and media outlets that fact-check and verify their content before publication should also be supported.

International partnerships and collaborations need to be established in order to combat deepfakes. Such partnerships can facilitate the sharing of knowledge, resources, and expertise, as well as provide a coordinated response to deepfake threats.

Finally, technology companies need to take responsibility for the content on their platforms. For example, they can invest in developing and deploying deepfake detection technologies and make it easier for users to report and remove deepfakes.

Conclusion

The looming spectre of deepfakes presents an unprecedented threat to national security. The rapid evolution and proliferation of this technology demands nothing short of a resolute and comprehensive response.

To safeguard democratic nations, governments must invest in cutting-edge deepfake detection technologies that can unmask digital imposters and expose malicious intent. Simultaneously, legal frameworks must be fortified, criminalizing the creation and dissemination of deepfakes while also providing robust protections for those whose reputations are vulnerable.

The battle against deepfakes cannot be won through technology and legislation alone. Citizens must be armed with the power of critical thinking and media literacy, thereby empowering them to discern truth from fabrication. By fostering a society that is professionally skeptical, informed, and resilient, governments can build a shield against the corrosive effects of deepfakes.

This fight transcends borders and requires global solidarity. Collaborative efforts between nations to share knowledge, resources, and expertise will be the cornerstone of defence. Together, an international alliance against deepfakes that is resilient and unwavering to preserving the integrity of collective societies can be forged.

Furthermore, technology companies must be held accountable. These companies wield immense power and influence, and as such, should prioritize the development and implementation of deepfake detection technologies in order to create user-friendly reporting mechanisms and swiftly remove deepfakes from their platforms. In doing so, technology companies can become defenders of national security.

Deepfakes pose a formidable challenge, but democratic governments and allies must stand united and vigilant against this challenge, as well as resolute in their commitments to protect democratic, diplomatic, and economic security.

Finding Signals in the Synthetic: Intelligence in the Era of Deepfakes

That deepfake videos will likely have a negative effect on our information environment has been well established. Less considered, however is the extent to which deepfakes will impact intelligence and national security agencies, including the threat environment they operate in, intelligence collection methods, the use of automated threat detection processes, the reception of intelligence products, and ethical dilemmas.

Despite the impressive technological nature of deepfakes, they are more likely to evolve current national security and intelligence threats rather than generate new ones. Instead, a more serious array of challenges may arise from ethical dilemmas that will require excellent judgement amid a collection of difficult choices ahead.

Evolving Threats

Generally, much of the literature on deepfakes focuses on the threats that they pose to democratic societies. Importantly, most of these threats are not likely to be new, but evolved and enhanced versions of threat-related activities that intelligence and national security agencies are already dealing with. This includes disinformation, the targeting of government/military personnel by adversarial forces, phishing/social engineering, and mimicking biometric data.

Disinformation

For the purpose of this paper, disinformation is defined as “false information that is intended to manipulate, cause damage, or guide people, organizations, and countries in the wrong direction”. Similar to disinformation, malinformation is information that stems from the truth but is often exaggerated in a way that misleads and causes potential harm[127]. Malinformation often stems from stolen or hacked information, some of which may be altered to lend credibility to a false narrative that an adversary wishes to emphasize, and then released on the internet for distribution.

Current online media ecosystems are awash with large amounts of disinformation and malinformation, which serve a number of different ends. From a national security perspective, this largely involves foreign interference and radicalization activities. It is well established that states such as Russia, China, and Iran engage in disinformation and malinformation campaigns to serve their political objectives. Likewise, violent extremist groups spread narratives about societal collapse, corrupt institutions, global conspiracy theories, and that violent, revolutionary action is required to restore humanity to its rightful condition (although this may be accomplished through the use of irony and memes)[128]. Importantly, despite their different political objectives, what often unites these two groups is their efforts to discredit and downplay democratic institutions, amplify conspiracy theories, and encourage distrust of what they see, generally, as “the system”.

The advantage of deepfakes for these actors is that they lower the cost of engaging in disinformation campaigns. Whereas in the past it may have taken time, effort, and skill to generate forgeries, and false information, deepfakes will quickly generate materials that can be utilized quickly, and spread worldwide even faster. Depending on how widespread and accessible deepfake tools are, they may also allow individuals to participate in information wars, thereby muddying an already complex information environment.

Targeting Government and Military Personnel

Adversarial actors will likely use deepfakes to target government, military, and national security personnel, usually for the purposes of making them targets or disrupting their work.

Already, there are disinformation campaigns taking place where western troops are posted, in order to breed mistrust and poor relations with civilian populations. For example, as Canadian troops deployed to Latvia as part of NATO’s Enhanced Forward Presence, a disinformation and malinformation campaign targeting them appeared[129]. These efforts at sowing distrust have been ongoing since 2017[130]. It is possible that deepfakes will be used to aid in these disinformation campaigns. Alternatively, the families of military personnel could be targeted with deepfakes involving their loved ones who may be serving abroad, in order to cause anguish and mental harm.

Social Engineering

A third area of concern is threats related to phishing or social engineering. Social engineering is the practice of obtaining confidential information by manipulation of legitimate users. Typically, social engineers use the telephone or internet to trick people into revealing sensitive information by pretending to be a figure of authority, co-worker, family member or even tech-support. This includes phishing, where a malicious actor sends an email mimicking or spoofing a specific, usually well-known brand, to convince someone to provide confidential information[131].

There is also concern that deepfakes, which can replicate the face, image, and voice of individuals, may trick people in more advanced ways. While this will almost certainly be a boon to criminals, adversarial intelligence agencies may use it to target politicians, intelligence officers or other holders of classified information in order to gain their trust and subsequent access to sensitive data. Moreover, in conflict, deepfakes may be used as ruses of war where fake videos or audio may be used to send false orders or commands to troops, or false information to disrupt military operations.

Hacking Biometric Data

Moving beyond social engineering to obtain classified or sensitive information, adversaries may use deepfakes to mimic biometric data in order to gain direct access. Research suggests that deepfakes may already have the capacity to fool biometric scanners, such as facial recognition systems[132]. Given that an increasing number of applications are collecting and using biometric data, it is likely that a significant number of institutions holding this data may either sell it or be susceptible to hacking[133]. Furthermore, this biometric data may be used to create deepfakes that are even more realistic.

Intelligence Collection

In responding to current, evolving, and future threats, national security and intelligence agencies are tasked with collecting information that pertains to their mandate. Unfortunately, it is likely that this too will be impacted by deepfakes in at least two ways: creating noise and targeting open-source information.

Creating Noise

Deepfakes may be employed as a disruption tool by adversarial states against intelligence collection. This could include signals intelligence should deepfakes be used to flood an information space, thereby creating lots of "noise" or false distraction. They may also be used tactically against a suspected, specific collection.

It is also possible that deepfakes could distort the perception of human sources who believe that an artificially generated conversation, video, or text is real, and subsequently pass that on to intelligence collectors in good faith. If a human source is unable to differentiate between true and fake information, it could impact intelligence collection and analysis.

Open-Source

A second issue relates to the use of open-source information by both government and non-government agencies. The 2022 invasion of Ukraine by Russia is the latest global event demonstrating the significance and value of open-source information and analysis[134]. Ranging from scraping social media through to analysis of publicly available satellite imagery, open-source techniques are being used to uncover troop movements, defensive fortifications, gain insight into the morale of combatants, verify attacks, losses, military strikes, and to investigate war crimes. Although the quality may vary, both national security agencies as well as journalists and humanitarian organizations have developed their own techniques or found reliable sources to inform their investigations. As such, open-source information is a prime target for deepfakes. Adversarial actors seeking to create division amongst allies, weaken resolve, deny war crimes, or falsify information will likely target open-source outlets with deepfakes. This may lead to incorrect reporting, which could then be used against open-source outlets to discredit their efforts. Even in a best-case scenario, deepfakes may make the already time-consuming job of open-source information verification much more difficult.

Automated Processes

It is also possible that deepfakes could impact automated processes designed to thwart adversarial activities. Data poisoning occurs when trawled data for deep-learning training of machine learning systems is compromised intentionally with malicious information[135]. Algorithms used to detect cyber-attacks, or disinformation/ malinformation campaigns could also be compromised through data poisoning of the large-scale sets of information they are trained on. Moreover, researchers have found that systems designed to detect deepfakes can be affected by data poisoning, rendering them less effective[136].

Reception of Information

As noted above, a key concern over deepfakes is the role they may play in worsening an already convoluted information environment. Therefore, while operating in an information space where the truth is increasingly contested, government officials, executives within national security and intelligence communities, and their analysts should anticipate challenges when it comes to having their findings accepted by the public or even politicians.

Intelligence assessments should always be questioned and/or interrogated by their audiences. However, where questioning is guided by accusations and challenges stemming from conspiracy theories, misinformation, disinformation and/or malinformation rather than the interests of good governance, the position of intelligence and national security agencies will be much more difficult. In particular, the social license that these agencies require to perform their jobs will be put at risk if a significant segment of the population rejects their findings outright, or ignores them because discerning the truth is seen as too difficult. This problem may be aggravated where these departments, and agencies have traditionally struggled with transparency.

Complicating matters further, warnings about deepfakes may actually reinforce the problem in some information ecosystems. Chesney and Citron note that efforts to warn the public about the pernicious effects of deepfakes may have a perverse outcome they call the "Liar's Dividend". In this scenario, individuals, corporations, and governments accused of engaging in harmful actions will be able to claim that any evidence produced, especially images, audio and/or videos are deepfakes, in an effort to dodge responsibility[137].

Ethics of Deepfakes

Given the concerns addressed above, researchers and scholars, particularly from a legal, scientific, and/or technical perspective, have focused on finding technical and regulatory solutions. Few articles have explored the ethical dilemmas generated by deepfakes, particularly for government departments and agencies. This paper will briefly discuss three of these dilemmas: the use of deepfakes and democratic norms, private sector dilemmas, and the risk of "over-hyping" the issue.

Should democracies use deepfakes?

The first challenge is that if deepfake techniques prove to be inexpensive and effective, there will be temptation to use them in the defence, security and intelligence operations of democratic countries. On the one hand, these states may wish to use these techniques because they are cost effective, and may be easier than other, riskier forms of intelligence gathering or covert activities.

For agencies that wish to use deepfakes, it may be argued that ruses of war have existed for centuries. Furthermore, a key goal of present information operations is the dissemination of propaganda in pursuit of a competitive advantage over an opponent[138]. This includes attempts to induce a sense of helplessness in an adversarial military or population, so they do not wish to fight[139]. Therefore, it will not be surprising if states manufacture deepfakes as a part of these campaigns to achieve their goals quickly, easily, and potentially with minimal bloodshed. Similarly, many intelligence agencies engage in disruption operations to prevent malicious activities from occurring on their territory or against their interests. Deepfakes could be used to mislead or fool adversaries with fake audio and video.

There is, however, a serious trade-off in doing so. It is expected that authoritarian states actively engage in propaganda and are very likely to turn to deepfakes to further their political objectives. However, given that democracies are grounded in the rule of law (however imperfect), they will not necessarily benefit in the same way from engaging in disinformation (nor is it clear if they are particularly good at information operations[140]). If it is known or believed that democracies, their militaries, and intelligence agencies are actively using deepfakes, the Liar's Dividend will certainly take effect in instances that will matter down the road, particularly if the West is trying to persuade new or skeptical audiences.

Moreover, as disinformation is widely recognized as a problem affecting democracies, it is questionable if creating more of it through deepfakes is a good idea. After all, western intelligence agencies seem to have had more luck with "pre-bunking" disinformation during the 2022 Russian invasion of Ukraine rather than creating an alternative set of lies[141].

Private Sector Dilemmas

A second series of ethical challenges is related to the role of the private sector in the creation, and detection of deepfakes. While AI and deepfake tools may enable a large number of independent, and proxy actors to engage in disinformation campaigns, it is possible that the real beneficiaries will be a small number of large, high net worth technology companies. Companies that have the means to amass, harness and process large datasets into machine learning systems, which can be used to both create, and detect deepfake content. How states work with these companies and use their products will require special care and consideration. Many machine learning datasets are based on images obtained through questionable means[142]. Concerns have been raised about racial bias in AI that can exacerbate systemic racism, and scientists have demonstrated that deepfake images can exacerbate racial bias in web-based face recognition APIs[143]. Deepfake algorithms may contain hidden racial and other biases that will affect outcomes.

Laws and privacy regulations will provide some guidance on what democratic states will be allowed to do. However, ethical judgement over what kinds of companies that states wish to engage with, how their practices are reviewed, and how to manage issues of accountability will be required.

Is the threat over-hyped?

Finally, for all the challenges that have been discussed in this paper, there is also a risk of exaggerating the threat. Disinformation is a serious problem, and even preliminary deepfakes may be contributing to it. However, many claims about the potential disruptive impact of AI-enabled propaganda are speculative and largely unscrutinised[144]. While AI will pose challenges, the present hype is not reality—deepfakes may be technically impressive, but this does not necessarily make their use practical. For example, a deepfake video of a world leader declaring war can quickly be checked and debunked simply by examining events on the ground.

Additionally, it is not immediately obvious that deepfake propaganda will be any more effective at sharing narratives than crudely-made images and memes, which are already widely and rapidly shared. Research has shown that fake news content spreads not because it is logical or realistic, but because it resonates emotionally with the sharer[145]. In this sense, states should be more concerned about certain narratives, rather than how good the content looks. States need to take deepfakes seriously—but in many cases they are evolving the current threat environment—not upending it. Therefore, overreaction to deepfakes may distort threat analysis and policy responses.

Conclusion

The above identifies some of the challenges (and opportunities) that national security, and intelligence agencies will face in the coming years. In doing so, it is argued that although the technology is impressive, deepfakes are more likely to evolve already existing threat-related activities, rather than generating new ones. If there is a silver lining to the deepfake dark cloud, it is that most democratic states are not starting from scratch but rather already have policies and procedures in place to help them manage deepfakes—although these too will need to evolve. For example, when collecting digital media, it will be important to establish chains of custody to help preserve and verify in the future or in law enforcement proceedings.

Many of the most challenging deepfake problems will not be solved with technology or law, but through ethical practices that will require good judgement. This includes thinking about how states should engage with the private sector, particularly those companies that already control large technology platforms, and what this means for oversight and review. Additionally, while there may be good reasons for democracies to consider the use of deepfakes for their own national security and intelligence operations, there may be more pitfalls than promise with this approach.

Developing Grounded, Human Rights-Centered Responses to Deepfakes, Synthetic Media, and Audiovisual Generative AI

Significant lessons can be learned from centering civilian and citizen concerns—including those of human rights defenders and journalists, globally—in the prioritization of risks, threats, and potential solutions in the areas of deepfakes, synthetic media, and audiovisual generative artificial intelligence (AI). Worldwide, these individuals and communities already confront harms similar to those emerging in the 'age of synthetic media'. However, despite being most at risk, these individuals and communities are marginalized from decisions on these new technologies.

Launched in 2018, WITNESS's 'Prepare, Don't Panic: Deepfakes and Synthetic Media' initiative has aimed to intervene early in the synthetic media ecosystem, focusing on technical infrastructure, emergent tools, digital literacy efforts, and policy and legislative aspects. The work is based on extensive research, industry consultation (including workshops in Europe, South Africa[146], Brazil[147], Southeast Asia[148], the United States[149]), and numerous online workshops and consultations across global geographies[150].

Threats and Risks from a Civil Society Perspective

Civil society actors, through the past five years of WITNESS consultations, consistently identify a set of existing harms and potential risks from synthetic media. Women are consistently identified as being particularly vulnerable to threats from synthetic media because of how the technology has enabled new forms of gender-based violence.

Synthetic media, or the existence of it, is used to exercise plausible deniability and dismiss demonstrably true evidence by claiming it is false (also known as the 'Liar's Dividend'), or to claim that all content cannot be trusted—often to discredit journalists, activists, and civil society organizations more broadly, along with the trustworthy content they put out to the world. Research participants expressed concern on the use of such claims (or fears) to justify the enactment of laws that limit speech more broadly.

The threat of synthetic media to spur misinformation and incite violence is consistently noted —particularly within existing vectors of rapid spread such as messaging apps—as well as how this can be used by foreign and domestic actors to target groups and communities who are already vulnerable based on ethnicity, religion, political identity, professional role(s) and/or other characteristics.

All of these threats are consistently connected to the existing challenges of media literacy, under-resourced journalistic capacity, and limited access to detection and authenticity tools for both critical civil society actors and individual citizens.

Workshop participants identified that these threat vectors combine with existing threat patterns directed towards civil society and citizens by their own governments in the context of closing civil society space. For example, spreading disinformation targeting civil society, surveilling and harassing of journalists and human rights defenders, and attempting the criminalization of their activities.

In the past year, as generative-AI based synthetic media tools have become more accessible, easier to use and more personalizable, more people have had the ability to engage with them. They have been able to imagine—or experience—how the tools could impact their lives. This shift has resulted in an emerging reevaluation of the risks and potential harms.

As people experimented with synthetic media tools and realized how easy they were to use to create individual content items, produce variants on content, and produce images of real-life events (with limited input data), the challenge of an information ecosystem flooded with synthetic content came up more regularly linked to the inadequacy of the volume of, for example, fact-checking responses. Participants placed these in critical contexts such as elections and public health crises[151].

Principles for Building Better Civilian Resiliency

In terms of building civilian resiliency to AI-based manipulation and synthesis, a number of core principles emerged through this research.

Prioritize: 1) people globally facing similar harms; and 2) journalists and civil society who support a reliable, trustworthy information ecosystem

A response to deepfakes and synthetic media, as well as the expansion in availability and ease of creation of audiovisual generative AI and deepfake technology requires attention to who is most at risk from both targeted attacks as well as a broader undermining of trust in critical content. In many cases, these same communities and stakeholder groups have already experienced related harms from previous technologies. For example, globally women journalists and public figures are targeted with non-consensual simulated sexual images, while human rights defenders and investigative journalists consistently face claims that their documentation and investigations have been falsified.

Similarly, these same constituencies already face real-world constraints on their capacity to respond. For instance, local journalists, local elections officials, and female and LGBTQI-identified community-level political figures often find themselves targeted, under-resourced, and over-burdened.

Avoid de-historicizing or decontextualizing synthetic media

Although synthetic media is an emerging technology, the threats it poses are not new. As noted above, existing experiences, particularly of vulnerable populations, critical civil society, and media intermediaries, should inform responses to the threats and opportunities of synthetic media. Marginalized populations know the ways in which they are targeted with disinformation, and community-based response strategies and fact-checkers have experience addressing more traditional methods of video and audio manipulation, or ‘shallowfakes’; while social media platforms are already grappling with how to handle satire (a common deepfake usage) on a global level.

Place firm responsibility across the pipeline of foundational model and tech builders, tech deployers, content creators and content distributors (media and social media)

Any solutions require careful attention across the pipeline of how synthetic media is made, from foundational model and tech builders, to the deployers of technology and content distributors. Responses should not place the burden or pressure on end-users to identify synthetic media or disclose their personal usage, in the absence of broader responsibility throughout the pipeline of how synthetic media is created.

For instance, it is not viable to double-down on a media literacy strategy focused on civilian forensic analysis of video. An example of this is the reliance on or promotion of tips that encourage someone encountering an image in their timeline to attempt to spot a potential visual glitch such as a distorted hand, created through the generative process. These tips rely on the current algorithmic 'Achilles Heel' and are often quickly remedied by technical progress.

Governments, social media platforms, technology companies, and news organizations all have a role in the development of mechanisms (such as regulation, policies, functionalities, processes, etc.) that proactively tackle threats without placing the responsibility on content creators or consumers alone, and locating responsibility (where relevant) with upstream stakeholders.

Developing a human rights-based technical infrastructure, norms, consistent global platform policies, and laws and regulations is key.

Governments and regulators can support a range of options at the technical and policy levels that help to establish clear human rights-based guardrails, mandate rights protections, and also pay close attention to critical rights issues around privacy and freedom of expression.

Actions to Support an Informed Digital Citizenry

These prioritized solutions reflect outcomes from WITNESS's research and require implementation that takes into account the factors identified above.

Media and digital literacy will not be enough, but are still as necessary as ever

Supporting media literacy more broadly is a critical component of societal and governmental responses. This is especially true considering that society is in the early stages of the use of synthetic media and deepfakes, and that so-called "shallowfakes"—where media is decontextualized, lightly edited or miscaptioned—are currently far more prevalent than synthetic content.

Techniques and approaches to synthetic media should not be developed in isolation from broader media literacy approaches or from approaches relevant to shallowfakes. For example, the SIFT approach[152] that focuses on the core principles of Stop, Investigate the source, Find alternative coverage, and Trace the original context, is an applicable methodology across broader media literacy, shallowfakes and emergent deepfakes. Media literacy campaigns need to be framed within the broader context of mis- and disinformation in an effort to promote critical consumption of content online. Media literacy campaigns should not focus on current technical flaws in particular generative techniques—or example, the well-known and now discredited tip that face-swap deepfakes do not blink—but on broader principles, and on the use of contextually appropriate detection and authenticity tools as they become available.

One key approach to consider in media literacy campaigns is to avoid adding to the existing hype around generative AI and synthetic media, and mitigating their impact in reducing the ability to trust content. Particularly for vulnerable communities and critical civil society voices, claims of '*it's a deepfake*' and the broader '*nothing can be trusted because anything can be falsified*' are already growing in prominence, and media literacy campaigns should be calibrated to the scale of the synthetic media problem and avoid alarmist rhetoric that compounds this issue.

In tandem with media literacy approaches aimed at the general population it is important to focus on 'the media's' literacy, and journalism's contribution to neither providing a simplistic educative response focused on visual tips, nor contributing to the panic cycle that is weaponized in some contexts against critical societal voices.

Detection tools should be accessible to those that need it most

Detection tools are part of a solution. In general, current detection tools are not reliable at scale, and require expert input to assess their results. In a number of global cases, the use by the general public of detection tools available online has contributed to confusion and increased doubt around real footage rather than contributing to clarity[153]. In the short-term, however, critical gaps exist in supporting better capacity and access to tools for journalists and fact-checkers, as they look to debunk realistic forgeries or dismiss claims that genuine journalistic multimedia content is fake.

Accessibility in these conditions extends beyond technical access to support in how to *effectively use* the detection tool(s), and to *effectively communicate* results to stakeholders.

While platforms can play a role in moderating synthetic media content, this should not be an automated process of unnuanced synthesis detection given the unreliability of existing detection models. The reality must be acknowledged that not only is most synthetic content personal communication or non-malicious/harmful, but that content will increasingly be a complex mix of synthetic and non-synthetic, and that detection efforts will have to accommodate for this. However, there is the potential for platforms to provide signals that could support both civil society and media entities doing analysis, as well as the broader media literacy noted above.

Verifiable provenance and watermarks can provide signals to support informed digital participation, but human rights and accessibility concerns should shape the infrastructure and tools

A range of initiatives are exploring how to provide authenticity and provenance signals to consumers of media (for example, initiatives such as the Coalition for Content Provenance and Authenticity, C2PA.org and the Content Authenticity Initiative, contentauthenticity.org). While there are a number of proposals for incorporating watermarking technologies into AI-generated technologies, all of these approaches rely on participation and integration across the pipeline of responsibility (outlined above) from foundational model developers, tool developers, social media platforms, and major news media outlets. They all hold a key responsibility to enable disclosure of how media encountered by citizens is created. Buy-in from social media platforms, major news media outlets and AI model and tool developers is essential as they play a key role in guaranteeing that the verifiable provenance or watermarks are part of, or attached to, the audiovisual content from early on in its lifecycle, but also in making sure that their users and viewers are effectively informed about the nature of the content they are consuming.

Within this context, a core responsibility for democratic government is to ensure that these technologies are deployed with privacy and access centered, and where necessary legislated. Work by WITNESS has identified a range of human rights concerns that are related to authenticity, provenance, watermarking and disclosure approaches—among the most prominent are ensuring that media provenance is not intrinsically or de facto connected to personal identity, and ensuring global stakeholder input[154]. As with synthetic media developers, the companies, organizations and governments behind initiatives promoting the use of provenance and watermarking technologies have a responsibility to assess these technologies for their potential to cause harm[155].

Synthetic media content moderation at scale still needs contextualized policies and local expertise to deal with threats from synthetic media and protect speech

Synthetic media will increase the amount of content created and shared online. Automating moderation, including by leveraging AI (e.g., detection tools, or provenance and watermarks), will be a facet of responses, however, these policies should be designed alongside impacted groups and based on human rights principles. In addition, there should be clear processes for local experts and experience to be involved in the moderation loop.

One key area of concern is to preserve options for the frequent satirical and parodic uses of synthetic media[156], while recognizing that this is a contested grey area and one where ‘gas lighting’ claims of humour are made around actual harmful or malicious content.

Democracy's New Challenge: Navigating the Era of Generative AI

Humanity is on the cusp of a new stage in human evolution, which will have a profound effect on society and democracy. One could call it the 'Era of Generative AI'—an epoch in which humanity's relationship with machines will change the very framework of society. Navigating this period of immense change, with both the opportunities and risks that it engenders, will be one of the biggest challenges for both democracy and society in this century.

Research into how the development of a new type of AI, so-called 'Generative AI', will impact humanity has been taking place for the last decade. The clue as to why this type of AI is so extraordinary is in its name: an emerging field of machine learning that allows machines to 'generate' or create new data or things that did not exist before.

The medium of this new data is any digital format. AI can create everything from synthetic audio to images, text and video. In its application, Generative AI can be conceived of as a turbo engine for all information and knowledge. AI will increasingly be used to not only create all digital content, but also as an automation layer to drive forward the production of all human intelligent and creative activity.

Picture a creative partner capable of writing riveting stories, composing enchanting music or designing breathtaking visual art. Now imagine this partner as an AI model—a tool that learns from the vast repository of digitized human knowledge, constantly refining its abilities to bring our most ambitious dreams to life.

This is Generative AI: a digital virtuoso that captures the nuances of human intelligence and applies this to create something new and awe-inspiring, or new and terrifying. Through tapping into the power of deep learning techniques and neural networks, Generative AI transcends traditional programming, effectively enabling machines to think, learn and adapt like never before.

This AI revolution is already becoming a fundamental feature of the digital ecosystem, seamlessly deployed into the physical and digital infrastructure of the internet, social media, and smartphones.

Nevertheless, while Generative AI has been in the realm of the possible for less than a decade, it was only last November that it hit the mainstream. The release of ChatGPT—a large language model (an AI system that can interpret and generate text) application—was an inflection point.

ChatGPT is now the most popular application of all time. It hit 100 million users within two months and currently averages over 100 million users per month. Almost everyone has a ChatGPT story, from the students using it to write their essays to the doctor using it to summarize patient notes. While there is huge excitement around Generative AI, it is simultaneously raising critical concerns around information integrity and brings into question our collective capacity to adapt to the pace of change.

From Deepfakes to Generative AI

The digital ecosystem that has been built over the past 30 years (underpinned by the internet, social media, and smartphones) has become an essential ecosystem for business, communication, geopolitics, and daily life. While the utopian dream of the Information Age has delivered, its darker underbelly has also become increasingly evident.

This ecosystem has empowered bad actors to engage in crime and political operations far more effectively and with impunity. Cybercrime is predicted to cost the world $10 trillion CAD in 2023. If it were measured as a country, then cybercrime would be the world's third-largest economy after the US and China.

However, it is not only malicious actors that cause harm in this ecosystem. The sheer volume of information to deal with, and humans' inability to interpret it, also have a dangerous effect. This is a phenomenon known as 'censorship through noise': when there is so much 'stuff' that one cannot distinguish or determine which messages one should be listening to.

All this was top of mind when encountering AI-generated content for the first time in 2017. As the possibility of using AI to create novel data became increasingly viable, enthusiasts started to use this technology to create ‘deepfakes’. A deepfake has come to mean an AI-generated piece of content that simulates someone saying or doing something they never did. Although fake, it looks and sounds authentic.

The ability for AI to clone people’s identities—but more importantly, to generate synthetic content across all forms of digital medium (video, audio, text, images)—is a revolutionary development. This is not merely about AI being used to make fake content—the implications are far more profound. In this new paradigm, AI will be used to power the production of all information.

Information Integrity and Existential Risk

In the 2020 book, *Deepfakes: The Coming Infocalypse,* it is argued that the advent of AI-generated content would pose serious and existential risks, not only to individuals and businesses, but to democracy itself. Indeed, in the three years since this book was published, humans have begun to encounter swathes of AI-generated content ‘in the wild’.

In early 2022, at the start of the Russian invasion of Ukraine, a deepfake video of Ukrainian President Zelensky, urging his army to surrender, emerged on social media. If this message had been released at a vitally important moment of the Ukrainian resistance, it could have been devastating. While the video was quickly debunked, this example of weaponized synthetic content is a harbinger of things to come.

Deepfake identity scams—like the one in which crypto scammers impersonated Tesla’s CEO Elon Musk—made more than $2.3 million (CAD) in six months in 2021, according to the Federal Trade Commission. Meanwhile, a new type of fraud (dubbed ‘Phantom Fraud’), in which scammers use deepfake identities to accrue debt and launder money, has already resulted in losses of roughly $4.5 billion (CAD).

Moreover, it is not only that every individual can become a victim of a deepfake attack. Cumulatively, the proliferation of AI-generated content has a profound effect on digital trust. Society was already struggling with the health of the information ecosystem before AI came into the equation—but what does it mean for democracy, and society, if everything humans consume online—the main diet feeding humans' brains—can be generated by artificial intelligence? How will humans know what to trust? How will they differentiate between authentic and synthetic content?

Safeguarding the integrity of the information ecosystem is a fundamental priority not only for democracy, but also for society as a whole. Not only can everything be 'faked' by AI, but the fact that AI can now 'synthesize' any digital content also means that authentic content (for example, a video documenting a human rights abuse or a politician accepting a bribe) can be decried as 'synthetic' or 'AI-generated'—a phenomenon known as 'the Liar's Dividend'.

The core risk to democracy is a future in which AI is used as an engine to power all information and knowledge—consequently degrading trust in the medium of digital information itself.

However, democracy (and society) cannot function if we cannot find a medium of information and communication that society can all agree to trust. It is therefore vital that society gets serious about information integrity, as AI becomes a core part of the information ecosystem.

Solutions: Authentication of Information

There are both technical and 'societal' ways in which to do this. One of the most promising approaches is the idea of content authentication. Rather than trying to 'detect' everything that is made by AI (which will be futile if AI drives all information creation in the future), the architecture of 'authentication' is embedded into the framework of the internet itself. This should be created with a cryptographic marker so its origin and mode of creation (whether it was made by AI or not) can always be verified. This kind of cryptography is embedded in the 'DNA' of the content, so it is more than just a watermark—it is baked in and cannot be removed or faked.

However, simply 'signing' this way is not enough. Society must also adopt an open standard to allow that 'DNA' or mark of authentication to be seen whenever humans engage with content across the internet—whether on email, YouTube, or social media. This open standard for media authentication is already being developed by the Coalition for Content Provenance and Authenticity (C2PA, C2PA.org) a non-profit organization focused on content provenance and authentication, which counts the BBC, Microsoft, Adobe and Intel among its members.

Ultimately, this approach is about radical transparency in information. Rather than adjudicating the 'truth' (a fool's errand), it is about allowing everyone to make their own trust decisions based on context. Just as one wants to see the label that indicates what went into the food one is eating, society needs to have the digital infrastructure in place that will allow its members to determine how to judge or trust online information that fuels almost every single decision individuals make.

Building Societal Resilience

However, the challenge cannot come down to technology alone. Society can build the tools for signing content, and the open standard to verify information across the internet, but the bigger challenge is understanding that humanity stands on the precipice of a very different world—one in which exponential technologies are going to change the very framework of society.

This means that old ways of thinking need to be updated. Our analogue systems are no longer fit for purpose. We must reconceptualize what it means to be a citizen of a vibrant democracy, and understanding is the first step. Ultimately, this is not a story about technology—this is a story about humanity.

While the recent advances in AI have kicked off much discussion about the advent of 'artificial general intelligence' (AGI)—the point at which machines take over as they become smarter than humans—society is not there (yet). Humans still have the agency to decide how AI is integrated into our society, and that is their responsibility. As a democracy, this challenge is one of the most important of our time—let us not squander our chance to get it right.

The Evolution of Disinformation: A Deepfake Future

An unclassified joint workshop of the Academic Outreach and Stakeholder Engagement (AOSE) and Analysis and Exploitation of Information Sources (AXIS) programs of the Canadian Security Intelligence Service (CSIS)

24 May 2023, Ottawa

AGENDA

8:30 - 8:45	Welcome and Opening prayer
8:45 - 9:00	Opening remarks
9:00 - 10:15	**Module 1 - Deepfakes in Context**
10:15 - 10:30	Break
10:30 - 12:00	**Module 2 - Capacities and Future Capabilities**
12:00 - 13:00	Lunch
13:00 - 14:30	**Module 3 - Implications for Intelligence and National Security**
14:30 - 14:45	Break
14:45 - 16:15	**Module 4 - Implications for Democracy**
16:15	Closing Remarks

Endnotes

1. A brief history of fake news (2022) Center for Information Technology and Society - UC Santa Barbara. Available at: https://www.cits.ucsb.edu/fake-news/brief-history.

2. Soll, J. (2016) The Long and Brutal History of Fake News. Available at: https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/.

3. Critical Disinformation Studies: History, Power, and Politics (2021). Available at: https://doi.org/10.37016/mr-2020-76.

4. Online disinformation (2023). Available at: https://www.canada.ca/en/campaign/online-disinformation.html.

5. United Nations (no date) Countering Disinformation | United Nations. Available at: https://www.un.org/en/countering-disinformation.

6. Public-Private Analysis Exchange Program (2021) "Increasing Threat of Deepfake Identities," https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf, 2021.

7. Johnson, D. and Johnson, A. (2023) "What are deepfakes? How fake AI-powered audio and video warps our perception of reality," Business Insider, 15 June. Available at: https://www.businessinsider.com/guides/tech/what-is-deepfake.

8. Sample, I. (2020) "What are deepfakes – and how can you spot them?," The Guardian, 13 January. Available at: https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them.

9. Public-Private Analysis Exchange Program (2021) "Increasing Threat of Deepfake Identities," https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf, 2021.

10. Treasury Board of Canada Secretariat (2019) Directive on Automated Decision-Making. Available at: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32745.

11. Copeland, B.J. (2023) *Artificial intelligence (AI) | Definition, Examples, Types, Applications, Companies, & Facts*. Available at: https://www.britannica.com/technology/artificial-intelligence.

12. Burns, E., Laskowski, N. and Tucci, L. (2023) "Artificial intelligence (AI)," *Enterprise AI*. Available at: https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence.

13. *Introducing ChatGPT* (no date). Available at: https://openai.com/blog/chatgpt.

14. Pocock, K. (2023) "What is ChatGPT and what is it used for?," *PC Guide*. Available at: https://www.pcguide.com/apps/what-is-chat-gpt/.

15. Allyn, B. (2022) "Surreal or too real? Breathtaking AI tool DALL-E takes its images to a bigger stage," *NPR*, 20 July. Available at: https://www.npr.org/2022/07/20/1112331013/dall-e-ai-art-beta-test.

16. Database, A.P.S. (no date) *AlphaFold Protein Structure Database*. Available at: https://alphafold.ebi.ac.uk/.

17. *AlphaFold* (no date). Available at: https://www.deepmind.com/research/highlighted-research/alphafold.

18. Bond, S. (2022) "As tech evolves, deepfakes will become even harder to spot," *NPR*, 3 July. Available at: https://www.npr.org/2022/07/03/1109607618/as-tech-evolves-deepfakes-will-become-even-harder-to-spot.

19. Tucker, P. (2021) "Deepfakes Are Getting Better, Easier to Make, and Cheaper," *Defense One*. Available at: https://www.defenseone.com/technology/2020/08/deepfakes-are-getting-better-easier-make-and-cheaper/167536/.

20. iProov (2022) "Deepfake Statistics & Solutions | Protect Against Deepfakes," iProov, 26 August. Available at: https://www.iproov.com/blog/deepfakes-statistics-solutions-biometric-protection.

21. CBC (2023) "Disinformation, Dictators & Democracy: A discussion with Maria Ressa and Ron Deibert," 10 May. Available at: https://www.cbc.ca/radio/ideas/disinformation-democracy-ressa-deibert-1.6837181.

22. The Nobel Peace Prize 2021 (10 December). Available at: https://www.nobelprize.org/prizes/peace/2021/ressa/lecture/.

23. Cole, S. (2023) "'You Feel So Violated': Streamer QTCinderella Is Speaking Out Against Deepfake Porn Harassment," Vice News, 13 February. Available at: https://www.vice.com/en/article/z34pq3/deepfake-qtcinderella-atrioc.

24. Farokhmanesh, M. (2023) "The Debate on Deepfake Porn Misses the Point," WIRED, 1 March. Available at: https://www.wired.com/story/deepfakes-twitch-streamers-qtcinderella-atrioc-pokimane/.

25. Gan, J. (2023) "Atrioc returns to Twitch six weeks after deepfake controversy, working on DMCA takedowns," Dexerto, 15 March. Available at: https://www.dexerto.com/twitch/atrioc-returns-to-twitch-six-weeks-after-deepfake-controversy-working-on-dmca-takedowns-2086445/.

26. Desk, I.T.W. (2018) "I was vomiting: Journalist Rana Ayyub reveals horrifying account of deepfake porn plot," India Today, 21 November. Available at: https://www.indiatoday.in/trending-news/story/journalist-rana-ayyub-deepfake-porn-1393423-2018-11-21.

27. Hany Farid, Robert Chesney, Danielle Citron (2020). "All's Clear for Deepfakes? Think Again". (11 May). Available at: https://www.ischool.berkeley.edu/news/2020/alls-clear-deepfakes-think-again.

28. Savin, J. (2022) "Deepfake porn is on the rise – and everyday women are the target," Cosmopolitan, 25 November. Available at: https://www.cosmopolitan.com/uk/reports/a41534567/what-are-deepfakes/.

29. Dunn, S. (2021) Women, Not Politicians, Are Targeted Most Often by Deepfake Videos. Available at: https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/.

30. Reuters (2021) "Fake Elon Musk giveaway featured in cryptocurrency scams-U.S. FTC," Reuters, 27 May. Available at: https://www.reuters.com/technology/fake-elon-musk-giveaway-featured-cryptocurrency-scams-us-ftc-2021-05-17/.

31. Serrano, J. (2022) "Please Don't Invest in This Crypto Scam Because Deepfake Elon Musk Told You To," Gizmodo, 27 May. Available at: https://gizmodo.com/elon-musk-deepfake-invest-bitcoin-scam-bitvex-1848982652.

32. Jenkinson, G. (2022) "'Yikes!' Elon Musk warns users against latest deepfake crypto scam," Cointelegraph, 26 May. Available at: https://cointelegraph.com/news/yikes-elon-musk-warns-users-against-latest-deepfake-crypto-scam.

33. Kohli, A. (2023) "From Scams to Music, AI Voice Cloning Is on the Rise," Time, 29 April. Available at: https://time.com/6275794/ai-voice-cloning-scams-music/.

34. Damiani, J. (2019) "A Voice Deepfake Was Used To Scam A CEO Out Of $243,000," Forbes, 3 September. Available at: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=120b6b6b2241.

35. Ropek, L. (2021) "Bank Robbers in the Middle East Reportedly 'Cloned' Someone's Voice to Assist with $35 Million Heist," Gizmodo, 30 October. Available at: https://gizmodo.com/bank-robbers-in-the-middle-east-reportedly-cloned-someo-1847863805.

36. Mileva, G. (2022) "The Ultimate Virtual Events Statistics You Need To Know in 2023," Influencer Marketing Hub [Preprint]. Available at: https://influencermarketinghub.com/virtual-event-statistics/.

37. Thubron, R. (2022) "FBI warns of more criminals using deepfakes in remote interviews for tech jobs," TechSpot, 29 June. Available at: https://www.techspot.com/news/95119-fbi-warns-more-criminals-using-deepfakes-remote-job.html.

38. Muncaster, P. (2022) FBI: Beware Deepfakes Used to Apply for Remote Jobs. Available at: https://www.infosecurity-magazine.com/news/fbi-beware-deepfakes-remote-jobs/.

39. Crews, J. (2023) "AI's Dark Side: The Potential Misuses of Artificial Intelligence & Why You Should Be Concerned," www.linkedin.com [Preprint]. Available at: https://www.linkedin.com/pulse/ais-dark-side-potential-misuses-artificial-why-you-should-crews.

40. "AI and Privacy: The privacy concerns surrounding AI, its potential impact on personal data," The Economic Times, 25 April 2023. Available at: https://economictimes.indiatimes.com/news/how-to/ai-and-privacy-the-privacy-concerns-surrounding-ai-its-potential-impact-on-personal-data/articleshow/99738234.cms?from=mdr.

41. Hern, A. and Milmo, D. (2023) "'I didn't give permission': Do AI's backers care about data law breaches?," The Guardian, 10 April. Available at: https://www.theguardian.com/technology/2023/apr/10/i-didnt-give-permission-do-ais-backers-care-about-data-law-breaches.

42. Hassan. S. (2023). "How AI Can Be Used to Manipulate People" Psychology Today, April 6 2023. Available at: https://www.psychologytoday.com/ca/blog/freedom-of-mind/202304/how-ai-can-be-used-to-manipulate-people#:~:text=By%20analyzing%20patterns%20in%20people%27s,punishments%20based%20on%20predicted%20behavior.

43. Rathenau Instituut (2022). "AI and manipulation on social and digital media". Available at: https://www.rathenau.nl/en/digitalisering/ai-and-manipulation-social-and-digital-media.

44. Marr, B. (2022) "The Problem With Biased AIs (and How To Make AI Better)," Forbes, 30 September. Available at: https://www.forbes.com/sites/bernardmarr/2022/09/30/the-problem-with-biased-ais-and-how-to-make-ai-better/?sh=8035ac547700.

45. Larkin, Zoe (2022). "AI Bias - What Is It and How to Avoid It?". Available at: https://levity.ai/blog/ai-bias-how-to-avoid.

46. Ruby, D. (2023) "Internet User Statistics In 2023 — (Global Data & Demographics)," DemandSage [Preprint]. Available at: https://www.demandsage.com/internet-user-statistics/#:~:text=5.07%20billion%20people%20around%20the,the%20internet%20as%20of%202023.

47. Statista (2023) Canada: number of internet users 2013-2023. Available at: https://www.statista.com/statistics/243808/number-of-internet-users-in-canada/#:~:text=Canada%3A%20number%20of%20internet%20users%202013%2D2023&text=As%20of%20January%202023%2C%20Canada,percent%20of%20the%20country%27s%20population.

48. Hancock, Jeffrey T. and Jeremy N. Bailenson (2021). "The Social Impact of Deepfakes". Cyberpsychology, Behavior, and Social Networking, Volume 24, Number 3, 2021. doi: 10.1089/cyber.2021.29208.jth.

49. Cook, J. (2022). "Deepfake Technology: Assessing Security Risk". Available at: https://www.american.edu/sis/centers/security-technology/deepfake_technology_assessing_security_risk.cfm#:~:text=Often%2C%20they%20inflict%20psychological%20harm,technology%20to%20conduct%20online%20fraud.

50. Pringle, E. (2023) "One of A.I.'s 3 'godfathers' says he has regrets over his life's work. 'You could say I feel lost,'" Fortune, 31 May. Available at: https://fortune.com/2023/05/31/godfather-of-ai-yoshua-bengio-feels-lost-regulation-calls/.

51. Metz, C. and Schmidt, G. (2023) "Elon Musk and Others Call for Pause on A.I., Citing 'Risks to Society,'" The New York Times, 29 March. Available at: https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.

52. Vallance, B.C. (2023) "Elon Musk among experts urging a halt to AI training," BBC News, 30 March. Available at: https://www.bbc.com/news/technology-65110030.

53. Brown, S. (2023). "Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI ". MIT Sloan. Available at: https://mitsloan.mit.edu/ideas-made-to-

matter/why-neural-net-pioneer-geoffrey-hinton-sounding-alarm-ai#:~:text=AI%20concerns%3A%20Manipulating%20humans%2C%20or%20even%20replacing%20them&text=And%20it%20seems%20very%20hard,own%20subgoals%2C%E2%80%9D%20Hinton%20said.

54. Goodyear, S. (2023) "The 'godfather of AI' says he's worried about 'the end of people,'" CBC, 4 May. Available at: https://www.cbc.ca/radio/asithappens/geoffrey-hinton-artificial-intelligence-advancement-concerns-1.6830857.

55. Taylor, J. and Hern, A. (2023) "'Godfather of AI' Geoffrey Hinton quits Google and warns over dangers of misinformation," The Guardian, 30 May. Available at: https://www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning.

56. Treasury Board Secretariat of Canada (2023b). "Directive on Automated Decision-Making". Available at: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592.

57. Canadian Heritage (2023). "Digital Citizen Initiative – Online disinformation and other online harms and threats". Available at: https://www.canada.ca/en/canadian-heritage/services/online-disinformation.html.

58. Reynolds, C. (2023) "AI pioneer Yoshua Bengio says regulation in Canada is too slow, warns of 'existential' threats," The Globe and Mail, 24 May. Available at: https://www.theglobeandmail.com/business/technology/article-ai-pioneer-yoshua-bengio-says-regulation-in-canada-is-too-slow-warns/.

59. Datta, B. (2017) "Can Government Keep Up with Artificial Intelligence?," NOVA | PBS, 10 August. Available at: https://www.pbs.org/wgbh/nova/article/ai-government-policy/.

60. Paas-Lang, C. (2023) "AI is having a moment. What should the government do about it?," CBC, 24 April. Available at: https://www.cbc.ca/news/politics/ai-regulation-mps-canada-1.6818095.

61. This chapter was derived from: Nieweglowska, M., Stellato, C., & Sloman, S. A. (2023). Deepfakes: Vehicles for Radicalization, Not Persuasion. Current Directions in Psychological Science, 09637214231161321.

62. Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019, September). *The State of Deepfakes: Landscape, Threats, and Impact.*

63. 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), 2056305120903408.

64. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

65. https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed.

66. Prior, M. (2013). Visual Political Knowledge: A Different Road to Competence? *The Journal of Politics*, 76(1), 41–57.

67. Messaris, P. &, Limus, A. 2001. The Role of Images in Framing News Stories. In Framing Public Life: Perspectives on Media and Our Understanding of the Social World, edited by Reese, Stephen D., Gandy, Oscar H., Grant, August E., 215–26. Mahwah: Lawrence Erlbaum.

68. Rim, S., Amit, E., Fujita, K., Trope, Y., Halbeisen, G., & Algom, D. (2015). How words transcend and pictures immerse: On the association between medium and level of construal. *Social Psychological and Personality Science, 6*(2), 123-130.

69. Kirkpatrick, A. (2020). The spread of fake science: Lexical concreteness, proximity, misinformation sharing, and the moderating role of subjective knowledge. Public Understanding of Science 30 (1), 55 - 74.

70. Barari, S., Lucas, C., & Munger, K. (2021). Political deepfake videos misinform the public, but no more than other fake media. OSF Preprints.

71. 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), 2056305120903408.

72. 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), 2056305120903408.

73. Lago F., Pasquini C., Böhme R. , Dumont H., Goffaux V. and Boato G., More Real Than Real: A Study on Human Visual Perception of Synthetic Faces [Applications Corner]. IEEE Signal Processing Magazine, vol. 39, no. 1, pp. 109-116, Jan. 2022.

74. Mori, M., K. F. MacDorman and N. Kageki. (2012). "The Uncanny Valley". IEEE Robotics and Automation Magazine, 19(2), 98–100.

75. Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. IScience, vol. 24 (11), 1–17.

76. Lee, J., & Shin, S. Y. (2022). Something that they never said: multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news. Media Psychology, 25(4), 531-546.

77. Pennycook, G., Epstein, Z., Mosleh, M. et al. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature* vol. 592, issue 7855 590–595.

78. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

79. Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019, September). *The State of Deepfakes: Landscape, Threats, and Impact.*

80. Chesney, R., and Citron, D.K. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. Calif. L. Rev. vol. 107, issue 6, 1753–1820.

81. Groh, M., Epstein, Z., Obradovich, N., Cebrian, M., & Rahwan, I. (2021). Human detection of machine-manipulated media. Commun. ACM 64, 10 (October), 40–47.

82. Yechiam, E., & Hochman, G. (2014). Loss attention in a dual task setting. Psychological Science, vol. 25, issue 2, 494-502.

83. Kirkpatrick, A. (2020). The spread of fake science: Lexical concreteness, proximity, misinformation sharing, and the moderating role of subjective knowledge. Public Understanding of Science 30, 55 - 74.

84. Bebbington, Jan & Russell, Shona & Thomson, I. (2017). Accounting and sustainable development: Reflections and propositions. Critical Perspectives on Accounting. 48. 10.1016/j.cpa.2017.06.002.

85. Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. Psychological Bulletin, 117(3), 497-529.

86. Tajfel, H. (1974). Social identity and intergroup behaviour. *Social science information*, *13*(2), 65-93.

87. Zhou, Y., & Shen, L. (2021). Confirmation Bias and the Persistence of Misinformation on Climate Change. Communication Research.

88. Liv., N & Greenbaum, D. (2020). Deep Fakes and Memory Malleability: False Memories in the Service of Fake News, AJOB Neuroscience, 11:2, 96-104.

89. Frenda, S. J., Knowles, E. D., Saletan, W., & Loftus, E. F. (2013). False memories of fabricated political events. Journal of Experimental Social Psychology, 49(2), 280-286. doi:10.1016/j.jesp.2012.10.013.

90. Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? The International Journal of Press/Politics, 26(1), 69–91.

91. Caramancion, K. M. (2021, April). The demographic profile most at risk of being disinformed. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-7). IEEE.

92. Harper, C. A., & Baguley, T. (2019). "You are fake news": Ideological (a)symmetries in perceptions of media legitimacy. OSF preprints.

93. Lawson, M. A., & Kakkar, H. (2022). Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news. Journal of Experimental Psychology: General, 151(5), 1154.

94. Caldwell, M., Andrews, J. T. A., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. *Crime Science, 9*(1), 14.

95. Europol. (2022). Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg. https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

96. Chesney, R., and Citron, D.K. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. Calif. L. Rev. vol. 107, issue 6, 1753–1820.

97. Liv., N & Greenbaum, D. (2020). Deep Fakes and Memory Malleability: False Memories in the Service of Fake News, AJOB Neuroscience, 11:2, 96-104.

98. Cialdini, R. B. (2018). Pre-suasion: A revolutionary way to influence and persuade. New York: Simon & Schuster Paperbacks.

99. Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research, 154*, 113368.

100. Harris, D. (2019, January 05). Deepfakes: False pornography is here and the law cannot protect You: Duke Law & Technology Review.

101. Gieseke, A. P. (2020). “The New Weapon of Choice”: Law’s Current Inability to Properly Address Deepfake Pornography. Vanderbilt Law Review, 73(5), 1479-1515.

102. 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), 2056305120903408.

103. Mitchell, Amy (2020, August 17). Many Americans SAY made-up news is a critical problem that needs to be fixed. Retrieved February 23, 2021, from https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/.

104. Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. Harvard Kennedy School (HKS) Misinformation Review.

105. Chesney, R., and Citron, D.K. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. Calif. L. Rev. vol. 107, issue 6, 1753–1820.

106. Barari, S., Lucas, C., & Munger, K. (2021). Political deepfake videos misinform the public, but no more than other fake media. OSF Preprints.

107. https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused#:~:text=People%20are%20arguing%20in%20court%20that%20real%20images%20are%20deepfakes%20%3A%20NPR&text=Press-,People%20are%20arguing%20in%20court%20that%20real%20images%20are%20deepfakes,claim%20that%20anything%20is%20fake.

108. Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research, 154*, 113368.

109. Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology*. General, 147(12), 1865–1880.

110. Ternovski, J., Kalla, J., & Aronow, P. M. (2021). Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments. OSF preprints.

111. Groh, M., Epstein, Z., Obradovich, N., Cebrian, M., & Rahwan, I. (2021). Human detection of machine-manipulated media. Commun. ACM 64, 10 (October), 40–47.

112. Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. Psychological Science in the Public Interest, 13(3), 106-131.

113. Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, 154, 113368.

114. Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? Business Horizons, 63(2), 135-146.

115. Chesney, R., and Citron, D.K. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. Calif. L. Rev. vol. 107, issue 6, 1753–1820.

116. Chesney, R., and Citron, D.K. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. Calif. L. Rev. vol. 107, issue 6, 1753–1820.

117. 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), 2056305120903408.

118. Mori, Masahiro (1970). "The Uncanny Valley". Translated by Karl F. MacDorman and Norri Kageki, 6 June 2012. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6213238&tag=1.

119. Cherry, K. (2022) "What Is the Uncanny Valley?" Verywell Mind. Available at: https://www.verywellmind.com/what-is-the-uncanny-valley-4846247.

120. Examples such as FakeApp or DeepFace Lab.

121. Examples include Lyrebird or Descript.

122. Examples include https://voice.ai/.

123. GAN-based deepfakes: These are the most advanced and sophisticated types of deepfakes. They use generative adversarial networks (GANs) to create realistic images, videos, or audio recordings. GAN-based deepfakes require significant computing power and technical expertise to create and are typically created by advanced AI researchers or professional teams. In other words, this is the type that may really hurt Canada's NATSEC interests.

124. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

125. https://abcnews.go.com/Politics/sharing-deepfake-pornography-illegal-america/story?id=99084399#:~:text=In%202019%2C%20synthetic%20media%20expert,%2C%E2%80%9D%20Ajder%20told%20ABC%20News.

126. BBC News, 'The fake video where Johnson and Corbyn endorse each other', 2019, accessed on 10 March 2022, https://www.bbc.com/news/av/technology-50381728.

127. Canadian Centre for Cyber Security, “How to identify misinformation, disinformation, and malinformation (ITSAP.00.300)”, February 2022. https://www.cyber.gc.ca/en/guidance/how-identify-misinformation-disinformation-and-malinformation-itsap00300.

128. See, for example, Whitney Phillips and Ryan M. Milner, *You Are Here: A Field Guide for Navigating Polarized Speech, Conspiracy Theories and Our Polluted Media Landscape*, Cambridge, MA: MIT Press, 2021.

129. Tom Blackwell, “Russian fake-news campaign against Canadian troops in Latvia includes propaganda about litter, luxury apartments”, *National Post*, 17 November 2017. https://nationalpost.com/news/canada/russian-fake-news-campaign-against-canadian-troops-in-latvia-includes-propaganda-about-litter-luxury-apartments.

130. Murray Brewster, “Canadian-led NATO battlegroup in Latvia targeted by pandemic disinformation campaign”, 25 May 2020. https://www.cbc.ca/news/politics/nato-latvia-battle-group-pandemic-covid-coronavirus-disinformation-russia-1.5581248.

131. Canadian Centre for Cyber Security, “Glossary”, Accessed 12 May 2023. https://www.cyber.gc.ca/en/glossary.

132. Hannah Smith, and Katherine Mansted, *Weaponised deep fakes; National security and democracy*. Canberra: Australian Strategic Policy Institute, 2020.

133. Ash Carter and Laura Manley, *Tech Factsheets for Policymakers: Deepfakes*, Cambridge, MA: President and Fellows of Harvard College, 2020; Office of the Privacy Commissioner, “Data at Your Fingertips Biometrics and the Challenges to Privacy”, February 2011. Also consider authoritarian states are collecting and centralizing biometric data as a part of their domestic surveillance programs. Riddle Russia, “Biometrics as a Kremlin tool”, 5 May 2023. https://ridl.io/biometrics-as-a-kremlin-tool/.

134. Economist, “Open-source intelligence is piercing the fog of war in Ukraine”, 13 January 2023. https://www.economist.com/interactive/international/2023/01/13/open-source-intelligence-is-piercing-the-fog-of-war-in-ukraine.

135. Payal Dhar, “Protecting AI Models from “Data Poisoning” New ways to thwart backdoor control of deep learning systems”, *IEEE Spectrum*, 24 March 2023 https://spectrum.ieee.org/ai-cybersecurity-data-poisoning.

136. Xiaoyu Cao, Neil Zhenqiang Gong, “Understanding the Security of Deepfake Detection”, in Pavel Gladyshev, et al. (eds) *Digital Forensics and Cyber Crime. ICDF2C 2021. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, Vol. 441. Springer, Cham, 2022. https://doi.org/10.1007/978-3-031-06365-7_22.

137. Chesney, R. and Citron, D.K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. SSRN Electronic Journal. doi:10.2139/ssrn.3213954.

138. RAND Corporation, “Information Operations”. https://www.rand.org/topics/information-operations.html Accessed 17 May 2023.

139. Clint Watts, *Messing with the Enemy: Surviving in a Social Media World of Hackers, Terrorists, Russians and Fake News*, New York: Harper Collins Publishers, 2018.

140. Brett Boudreau, “The Rise and Fall of Military Strategic Communications at National Defence 2015-2021: A Cautionary Tale for Canada and NATO, and a Roadmap for Reform”, Canadian Global Affairs Institute, May 2022. https://www.cgai.ca/the_rise_and_fall_of_military_strategic_communications_at_national_defence_2015_2021.

141. Stephanie Carvin, “Deterrence, Disruption and Declassification: Intelligence in the Ukraine Conflict”, CIGI Online, 2 May 2022. https://www.cigionline.org/articles/deterrence-disruption-and-declassification-intelligence-in-the-ukraine-conflict/ ; David Klepper, ‘”Pre-bunking’ shows promise in fight against misinformation”, Associated Press, 24 August 2022. https://apnews.com/article/technology-misinformation-eastern-europe-902f436e3a6507e8b2a223e09a22e969 .

142. Chloe Xiang, “AI Is Probably Using Your Images and It’s Not Easy to Opt Out”, *Vice*, 26 September 2022. https://www.vice.com/en/article/3ad58k/ai-is-probably-using-your-images-and-its-not-easy-to-opt-out.

143. Shahroz Tariq, Sowon Jeon and Simon S. Woo, “Evaluating Trustworthiness and Racial Bias in Face Recognition APIs Using Deepfakes”, Computer, Vol. 56, No. 5, May 2023. doi: 10.1109/MC.2023.3234978.

144. James R. Ostrowski, “Shallowfakes: The danger of exaggerating the AI disinfo threat”, *The New Atlantis*, Spring 2023. https://www.thenewatlantis.com/publications/shallowfakes.

145. Cameron Martel, Gordon Pennycook and David G. Rand, “Reliance on emotion promotes belief in fake news”, *Cognitive Research: Principles and Implications,* Vol. 5, Article 47, 2020. https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-020-00252-3.

146. What We Learned from the Pretoria Deepfakes Workshop (Full Report) - WITNESS Blog. Available: https://blog.witness.org/2020/02/report-pretoria-deepfakes-workshop/.

147. WITNESS Media Lab | Deepfakes: Prepare Now (Perspectives from Brazil) - WITNESS Media Lab. Available: https://lab.witness.org/brazil-deepfakes-prepare-now/.

148. WITNESS Media Lab | Deepfakes: Prepare Now (Perspectives from South and Southeast Asia) - WITNESS Media Lab. Available: https://lab.witness.org/asia-deepfakes-prepare-now/.

149. Deepfakes and Disinformation – MediaJustice. Available: https:/mediajustice.org/news/deepfakes-and-disinformation/.

150. For reports wit.to/Synthetic-Media-Deepfakes.

151. WITNESS, Fortifying the Truth in the Age of Synthetic Media and Generative AI, May 2023, https://blog.witness.org/2023/05/generative-ai-africa/.

152. See https://hapgood.us/2019/06/19/sift-the-four-moves/ and https://ctrl-f.ca/.

153. https://blog.witness.org/2021/07/deepfake-detection-skills-tools-access/.

154. Ticks or It Didn't Happen: Confronting Key Dilemmas in Authenticity Infrastructure for Multimedia https://lab.witness.org/ticks-or-it-didnt-happen/ and 'Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism' https://journals.sagepub.com/doi/10.1177/14648849211060644 .

155. https://blog.witness.org/2021/12/witness-and-the-c2pa-harms-and-misuse-assessment-process/.

156. Just Joking: Deepfakes, Satire, and the Politics of Synthetic Media, https://cocreationstudio.mit.edu/just-joking/.