TS//SI//REL TO CAN, FVEY

Communications Security Establishment | Centre de la sécurité des télécommunications

CANADIAN CENTRE FOR CYBER SECURITY

(S) Deepfakes and Disinformation: The Malicious Use of Machine Learning Enabled Technology

Canada

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

(U) ABOUT THIS DOCUMENT

(U) This assessment is intended to inform Canadian decision-makers about Machine Learning-Enabled (MLE) technologies and their use by cyber threat actors. MLE technologies offer advanced disinformation capabilities in support of online foreign influence activities (OFIA). MLE technologies are tools that can generate synthetic media (text, images, video, etc.), mimic user behaviours, and manipulate algorithms in order to advance and amplify disinformation. This assessment also provides an overview of the efficacy of several publicly available detection models used to identify machine-generated media.

(U) Limitations

(U) This assessment is informed by data science analysis of synthetic detection methods. Twitter was selected to evaluate the efficacy of multiple detection models due to the availability of publicly available datasets of alleged state-sponsored foreign influence activity on the platform, periodically released by the company since 2019. While detection models were used to evaluate the presence of synthetic text, images, and videos, machine-generated audio detection was not included in the scope of this assessment.

(U) Key Terms

(U) Machine learning (ML) is a field of research into methods that allow machines to learn how to complete a task from given data without explicitly programing a step-by-step solution. ML programs can often approach or exceed human performance; as such, machine learning is considered a sub-discipline of Artificial Intelligence (AI) research.

(U) "Synthetic content" and **"synthetic media"** refer to content that is machine-generated with little to no human assistance. "Deepfakes" are a subset of synthetic media limited to ML-based image and audio synthesis techniques. A "synthetic image" refers to an image generated without a reference image using Machine-Learning Enabled (MLE) technology. In contrast, manipulated images (such as those generated by using software such as Photoshop) begin with a reference image as its base which is then modified to create a new image.

(U) Foreign influence activity occurs when foreign actors covertly create, disseminate or amplify misinformation or disinformation to influence the beliefs or behaviours of the citizens of another state. OFIA occurs when foreign actors covertly create, disseminate or amplify misinformation or disinformation to influence the beliefs or behaviours of the citizens of another state.

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

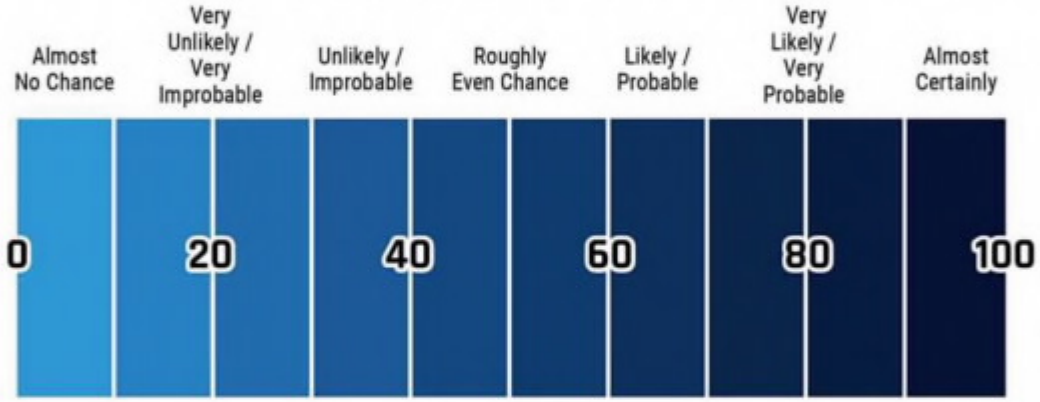
(U) Assessment Process

(U) This assessment is based on an analytical process that includes evaluating the quality of available information, exploring alternative explanations, mitigating biases, and using probabilistic language. We use the terms "we assess" or "we judge" to convey an analytic assessment. We use qualifiers such as "possibly," "likely" and "very likely" to convey probability.

(U) This assessment is based on information available as of 16 March 2022.

(U) The Canadian Security Intelligence Service (CSIS) and Privy Council's Office (PCO) reviewed and provided comments on this report.

The chart below matches estimative language with approximate percentages. These percentages are not derived via statistical analysis, but are based on logic, available information, prior judgements, and methods that increase the accuracy of estimates.



(U) For questions or feedback please contact us via your SECRET network:

(S//REL TO CAN, FVEY) To contact us via your TOP SECRET network:

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

(U) KEY JUDGEMENTS

- (TS//SI//REL TO CAN, FVEY) We assess that state-sponsored cyber threat actors will almost certainly increase their use of Machine Learning Enabled (MLE) technologies to enhance their online foreign influence activities (OFIA)
- (TS//SI//REL TO CAN, FVEY) States with sophisticated capabilities, such as Russia and the People's Republic of China (PRC) are very likely to generate deepfakes and generative adversarial networks (GAN) fake images that are harder to detect
- (TS//SI//REL TO CAN, FVEY We judge that it is likely that foreign adversaries will develop authentication methods to counteract the effects of
- (S//REL TO CAN, FVEY) We also judge that cybercriminals of varying sophistication will almost certainly increase their use of MLE technologies
- (S//REL TO CAN, FVEY) We assess that, given the current ineffectiveness of synthetic content detection models and the increasing availability of the MLE technologies that generate them, it is likely that state-sponsored OFIA aiming to propagate disinformation

(U) INTRODUCTION

(U) Since 2016, Machine Learning-Enabled (MLE) technologies that can generate fake text, images, audio, and video documents have become increasingly accessible to a range of cyber threat actors, including state-sponsored actors. We judge that Canadians have very likely been exposed to synthetic content circulating on social media.[1] Consequently, MLE technologies represent a growing and evolving threat to Canada's information ecosystem, including its media and telecommunication landscape and the structures in which information is created, shared, and transformed. MLE technologies can generate convincing synthetic content which can be used to augment disinformation campaigns, covertly manipulate online information, and influence opinions and behaviours.

(U) This assessment:

- (U) Explains how some MLE technologies can create realistic-looking content and how cyber threat actors, including state-sponsored actors, use this content to spread disinformation.
- (U) Evaluates the effectiveness of some widely-available detection models at identifying synthetic content on social media.
- (U) Assesses the current and future threat of MLE technologies on the Canadian information ecosystem.

(U) Enhancing Online Foreign Influence Activity with Machine Learning

(TS//SI//REL TO CAN, FVEY the PRC, Iran, and Russia, are investing in the research and development of MLE technologies and are adapting them to create synthetic content for the purposes of conducting OFIA.[2]

TS//SI//REL TO CAN, FVEY

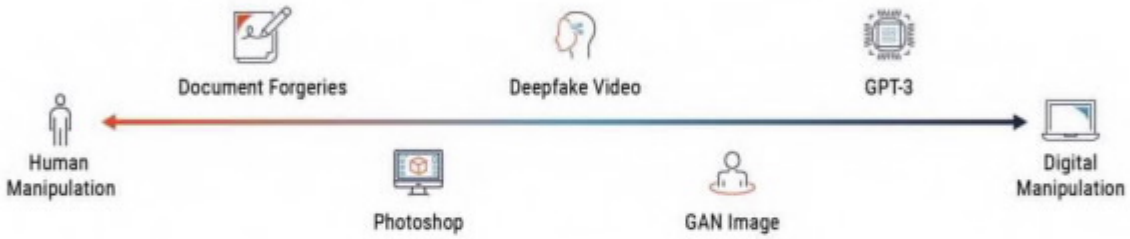
CANADIAN CENTRE FOR CYBER SECURITY

Further, we assess that some state actors are very likely to increase their use of MLE generated synthetic content in ongoing influence operations, targeting public opinion on popular social media platforms.[3]

(U) We assess that Canada may be particularly vulnerable due to the high intake of social media content by Canadians. In January 2021, the estimated number of Canadian-owned accounts on social media platforms Facebook, Instagram, Twitter, TikTok, WeChat and YouTube totalled 67.1 million.[4] In 2019, almost 50% of Canadians aged between 18 and 24 relied on social media as their main source of news.[5]

(U) Synthetic content is defined as a broad spectrum of generated or manipulated digital content, including images, video, audio, and text, that is often meant to deceive the recipient. It has evolved over time from relying heavily on human manipulation and person-directed techniques to becoming machine-generated (see Figure 1).[6]

(U) Figure 1. Synthetic Content's Shift toward Digital Manipulation



(U) MLE technologies that generate synthetic content using neural networks (see text box) have progressed to the point where the content they produce is often nearly indistinguishable from legitimate products.[7] While numerous MLE technologies exist today, the following applications are widely used in both legitimate and malicious activities:

- (U) MLE text generators, such as GPT-2 and GPT-3, that can be used to write synthetic text about a particular topic in a particular style.
- (U) MLE image generators like GAN Lab or Mimicry that can fabricate fake images that are almost indistinguishable from real ones.
- (U) MLE video and audio synthesis tools, like DeepFaceLab and Lyrebird, that can create fake videos, often referred to as deepfakes.

(U) What are Neural Networks?

(U) Artificial neural networks are flexible models that can be trained to perform and automate very specific and complex tasks, such as generating realistic videos of events that never occurred (commonly referred to as deepfakes). They are able to identify and learn the relationships/patterns that exist within extremely large datasets and build up complex representations of this data. This makes them an essential component of MLE technologies that can generate convincing synthetic media.

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

(U) TEXT GENERATORS & ONLINE FOREIGN INFLUENCE CAMPAIGNS

(U) In August 2019, OpenAI, an artificial intelligence (AI) research and deployment company, released a partial version of its Generative Pre-trained Transformer 2 (GPT-2), a language model capable of generating paragraphs of coherent text that are virtually indistinguishable from human writing.[10] OpenAI initially released an extremely restricted version of the system due to concerns about its possible applications for "reducing the cost of generating fake content and waging disinformation campaigns".[11] This is because GPT-2 has the capacity to manufacture believably-human text on any number of topics at an unprecedented scale.[12] GPT-2 was shown by researchers to have the potential of being weaponized to generate convincing extremist content.[13]

(U) OpenAI: A Key AI Developer

(U) OpenAI is a leader in developing generative pre-trained transformer models, the main technical component in its GPT-2 text generator. The San Francisco-based company was founded in 2015 as a non-profit and within five years it became one of the leading AI research labs in the world. In July 2019, Microsoft invested $1 billion in OpenAI to get exclusive access to its GPT-3 text generator source code.[8] The company is currently developing other machine learning products including an MLE system that translates natural language into code and a neural network that can create images from text captions.[9]

(U) The full version of GPT-2 has been publicly available since November 2019.[14] Despite concerns about GPT-2's malicious applications, OpenAI continued its development and, in June 2020, released its newest version, GPT-3, once again limiting access to the model citing concerns over potential misuse.[15] Several other AI developers have also released different versions of MLE language generators, most of which are also publicly accessible.

(U) Malicious use of MLE Text Generators

hostile state actors, are developing these types of text generating capabilities MLE text generators can help solve most of these problems.[17] MLE text generators can also For example, researchers found that GPT-3's text generating capabilities combined with other AI-based personality analysis tools can be used to conduct mass phishing campaigns that effectively target every single recipient with a personalized email.[18]

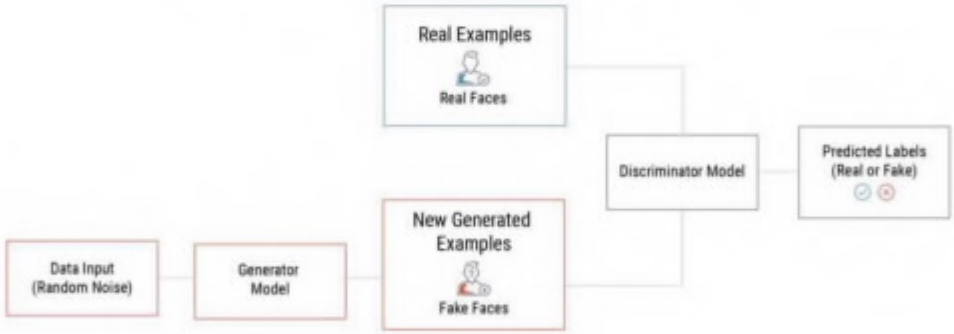
(TS//SI//REL TO CAN, FVEY) In recent years, the PRC and Russia have moved to research and deploy MLE text generators:

CANADIAN CENTRE FOR CYBER SECURITY

(U) FAKE PEOPLE: CREATING FALSE PERSONAS WITH GAN TECHNOLOGY

(U) The arrival of Generative Adversarial Network (GAN) technology in 2014 marked the start of sophisticated synthetic image creation. Prior to GANs, neural networks could create images, but these were unconvincing, blurry or missing key elements. Synthetic images are currently nearly indiscernible from real images.[27] What distinguishes GAN generated images from previous image creation/manipulation tools is the way that the model is trained (see Figure 2). If a user wishes to create synthetic images of human faces, they begin by inputting original images of human faces (labelled "real faces"). The GAN's generator model, which is a neural network, then creates synthetic images of human faces based on patterns it has found across the input dataset (i.e. the original images labeled "real faces"). The GAN then uses a discriminator model, a different neural network, to test if the synthetic images could pass as being labeled "real faces". Continuous training of the generator and discriminator models improves the synthetic images until they become nearly indistinguishable from images of real human faces.[28]

(U) Figure 2. Training a Generative Adversarial Network



(U) Malicious uses of GAN Generated Images

(U) We assess that state-sponsored cyber threat actors are highly likely to continue to use GANs to manufacture synthetic images for fake social media accounts that post or amplify disinformation and propaganda and reinforce online influence operations. We also assess that state-sponsored cyber threat actors will almost certainly use these techniques to create

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

personas for cyberespionage operations. A profile picture can make an account more believably human. Unlike stolen or purchased pictures of real people, GAN generated images cannot be traced using basic techniques such as a reverse image search.

- (U) In February 2021, Russian media reports noted that the Russia government used GAN-generated faces to create Instagram accounts used to geotag a false location for a real anti-Kremlin protest, ostensibly as a ruse to attract Alexei Nalvany supporters to a heavily-policed area where they would likely be stopped and/or arrested.[35]
- (U) Russia has long been suspected of using GAN images to create false personas meant to lure its adversaries into divulging information relevant to its espionage operations. In June 2019, Russian intelligence services were suspected of creating the "Katie Jones" online persona, who claimed to be a "Russia and Eurasia fellow" at a think tank in Washington D.C. and was able to add dozens of top US government officials to her professional network on LinkedIn.[36]

(U) SEEING IS BELIEVING: DEEPFAKE VIDEOS & FACE SWAPS

(U) The term "deepfake", combining "machine deep learning" and "fake", uses MLE image and audio synthesis techniques to generate fake videos that appears realistic and genuine to viewers. Deepfake video creators can use these techniques to superimpose a target person's features, expressions and movements onto another person's face (see Figure 3).

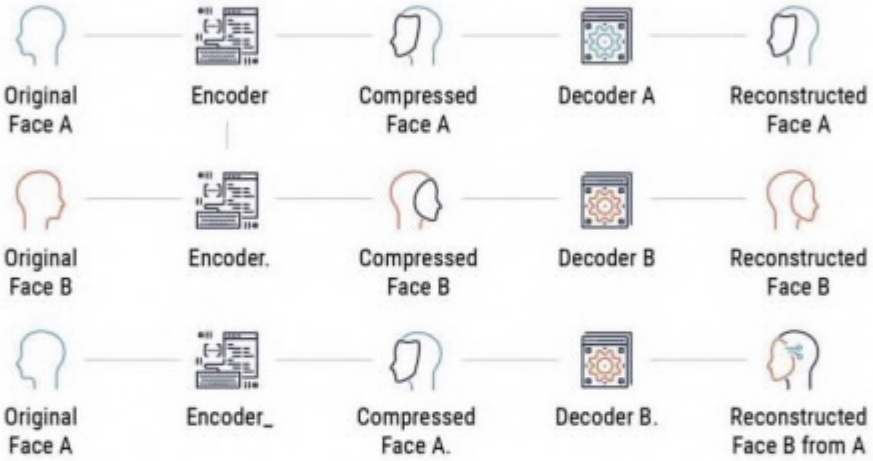
(U) Figure 3. Examples of Deepfakes[38]

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

(U) There are a number of ways in which a deepfake video can be generated using neural networks. One technique is called "face swapping" (shown in Figure 4). The first step is to run thousands of pictures of two people into an algorithm called an encoder. The encoder identifies the key features of a face (Face A) and compresses the image. The same encoder then repeats this process for the second face (Face B). The compressed image associated with face A and B are then fed into two different decoder algorithms. The first decoder is trained to recover the first person's face (A), and the second decoder is used to recover the second person's face (B). To construct an image of person B with the form, expression, and orientation of person A, the compressed image for person A is fed into the decoder for person B.[39]

(U) Figure 4. How Encoder and Decoder Algorithms Create Deepfake Videos



(U) Once the deepfake model is trained on enough data, it can reconstruct a series of synthetic images in order to make a fake video of a targeted person.[40] And, by reverse engineering real audio-video recordings, MLE technologies can generate deepfake videos that convincingly mimics an individual's visual and auditory style of speech.[41]

(U) Malicious use of Deepfake Videos

MLE technologies enhance disinformation operations and becoming more ubiquitous. Already, freely-available deepfake applications, offer open source code to create crude deepfake videos deceptive and effective at causing reputational damage.

(TS//SI//REL TO CAN, FVEY) State-sponsored cyber threat actors developing deepfake video and audio capabilities in order to supplement their influence campaigns.[42] Deepfakes can be used to generate propaganda, create false-flags or compromising material, and fuel conspiracy theories. For example, .[43] Deepfake

recordings of a political candidate delivering a controversial speech or in an embarrassing or compromising situation could also be created to help recruit terrorists, discredit rivals or, in a democratic situation, influence voters.[44] There is already a growing collection of deepfake videos circulating online that show high-profile leaders such as Donald Trump, Barack Obama, and Vladimir Putin making false statements.[45]

(TS//SI//REL TO CAN, FVEY We judge that even if deepfakes do not succeed in deceiving viewers, they can still cause severe reputational damage and distress to victims. Deepfake detection cannot always rectify damage caused by a deepfake video, especially when it depicts embarrassing and explicit content. Conversely, widespread awareness about deepfake technology will likely lead to increased skepticism of media, causing some to doubt or disregard truthful information.[46] This distrust can negatively impact public discourse and greatly exasperate political unrest.[47]

foreign adversaries will develop authentication methods to counteract the effects

(TS//SI//REL TO CAN, FVEY)

- (U//OUO) Suspected Russian state-sponsored cyber threat actors have reportedly used deepfake videos to impersonate a person during online video conferencing calls.[50] On 23 April 2021, a Latvian television news program aired a deepfake video of Russian opposition leader Leonid Volkov. The deepfake video depicted Mr. Volkov giving an interview over video conference and a deepfake video of Volkov was later used in video conferencing calls with Ukrainian and British politicians. The creators of the video—Vladimir Kuznetsov and Alexei Stolyarov—are accused of working with Russian security services to develop their deepfake capabilities as a part of a broader pro-Russian influence campaign against Western governments, human rights organizations, and Russian political opposition.[51]
- (U//OUO) Between 2015 to 2020, over $125 million USD was transferred to fraudsters and cybercriminal organisations using voice deepfakes in voice phishing scams or other voice-based social engineering scams. Voice deepfakes can be used to imitate the voice, tonality and punctuation in order to impersonate someone over a call.[52]
-

(U) We assess that it is highly likely that deepfakes will increasingly be used to blackmail individuals, chiefly through 'sextortion' campaigns, using the threat of reputational damage to extract payment. Between August 2019 and January 2021, third party monitoring recorded a drastic uptick in Dark Web source activities on deepfake-related topics, particularly the creation of sexually explicit deepfakes, as well as an increase in advertising for customized deepfake service offerings.[56] Simply by having access a few original pictures, cybercriminals can threaten to send deepfake pornographic content to all the victims contacts.[57]

(U) Deepfake's Disproportionate Impact on Women

(U) We assess almost certainly that the most immediate illicit use of deepfake technologies has been directed at women who are at a higher risk of being non-consensually depicted in sexually explicit synthetic content. Some researchers estimate that 95% of all deepfake videos on the Internet contain non-consensual pornography and that about 90% of these depict women.[54] Some of the most popular MLE tools available today are widely available apps that "digitally undress" pictures and generate personalized deepfake pornographic material.[55]

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

(U) DIFFICULT TO DETECT: THE RISING THREAT OF MLE TECHNOLOGY

(U) We assess it is almost certain that the amount of synthetic content circulating on social media is increasing as MLE technologies become more accessible. In September 2019, research firm DeepTrace Labs found 14, 678 deepfake videos posted online, representing an almost 100% increase since December 2018.[58]

(S//REL TO [60] Within the context of synthetic imaging and deepfake video, this could provide them with an advantage since large databases like these could be used to generate new synthetic content that is harder to detect. We assess that it is very likely that these state cyber threat actors will use this data to generate deepfakes and GAN fake images that are harder to detect

(U//OUO) We assess that current publicly available detection models are very likely ineffective in identifying deepfakes in uncontrolled environments, meaning non-test environments in which the ratio deepfake versus real content is not known (e.g., social media platforms' content). This ineffectiveness combined with the overall increasing availability of MLE technologies leads us to conclude that synthetic content detection methods will struggle to keep up with MLE technologies as they continue to improve. We assess that, given the current ineffectiveness of synthetic content detection models and the increasing availability of the MLE technologies that generate them, it is likely that state-sponsored OFIA aiming to propagate disinformation will increasingly go undetected, appear authentic or be produced on a mass scale rendering manual identification impossible. We also assess that it is very likely that as technology develops and becomes harder for humans to detect, it will also be better at fooling detection models, impacting social media companies' ability to detect and remove synthetic content.

(U) OUTLOOK

(TS//SI//REL TO Meanwhile, the adoption of MLE technology these technologies will very likely play an increasingly important part in their strategic influence operations [62] We assess that this will enhance their ability to create disinformation and manipulate online public discourse, potentially harming Canada's information ecosystem.

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

(TS/ Synthetic content produced by MLE technologies is already in circulation on social media platforms frequented by Canadians. [63] We also deem it likely that, within the next year, Canadians will be increasingly targeted by cybercriminals using MLE technologies, particularly deepfakes for blackmail, extortion, fraud or other money-making scams.

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

(U//OUO) Annex A: Detection Model Performance Evaluation

Methodology	
	In order to evaluate the performance of some currently available detection models we tested them on tweets from Twitter's publicly available disclosure of state linked information operations. We developed and used detection models based on code found on publicly accessible code sharing websites and preprint servers. These detector models are almost certainly different than those used by social media companies such as Meta (Facebook and Instagram) which are not publicly available. Our aim was to assess their ability to identify synthetic content on social media platforms like Twitter in a novel way. Detection model performance is usually evaluated by academic researchers using controlled datasets where fake and real posts are known. In contrast, raw Internet data, like Tweets, tend to be of mixed quality and none are labelled as real or fake. As such, our research tested the models' performance in the "uncontrolled environment" of the Twitter dataset.
Key Findings	
	We judge that the accuracy of the detection models dropped significantly when these were tested on uncontrolled environments like Twitter. The overreliance on the "fake" label was a reoccurring issue across all synthetic content detection models, leading to multiple false positives (real content labelled as synthetic).
	We assess that it is likely that publicly available MLE text generator detection models are ineffective in detecting text generated by GPT-2 on social media, primarily because the text is short. For example, when we tested the detection accuracy of a GPT-2 detection model on Tweets we found that the model struggled to properly label Tweets, particularly those which contained porn or spam. When we tested the detection model on Tweets from prior to GPT-2's release in 2019, we found that the detection model flagged roughly the same amount of computer generated text, meaning that a majority of the Tweets labeled as fakes were in fact false positives.
	We also judge that the currently available detection models for both GAN images and deepfake videos are likely ineffective at detecting synthetic content based on the propensity of a sample model we tested to label almost all content found on social media to be fake.
Comparative Analysis	
	By comparing multiple detection models' performance in an uncontrolled environment like Twitter we were able to assess that deepfake video is likely the hardest to detect. Since videos can be modified in part, it is extremely hard for detectors to identify which part of the video is a deepfake and which is not. Computer-generated text is the next hardest to detect due to the fact that most

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

	Internet speech is short (i.e. tweets) and it is very difficult for detector models to differentiate between human and computer-generated text only using a few words. Of the three types of synthetic content in this assessment, GAN images are easiest to detect, however, detector models can still be confused by low quality images and by faces that are out of focus. Moreover, based on how labour intensive each type of MLE generated synthetic content is we estimate that generated text is easier to make, followed by GAN imaging and finally deepfake video.

(U//OUO) Annex B: Detection Model Accuracy

GPT-2	
Results of detection accuracy in controlled environment	A version of a model known as RoBERTa, which was trained to detect generated text from GPT-2, achieved 96% accuracy in a controlled environment, with most errors coming from false positives (real text mistaken for being computer generated "deepfake" text). The length of the text also impacted detection performance, only achieving a 92% accuracy in detection of the shortest 10% of texts.
Results of detection accuracy in uncontrolled environments	When RoBERTa was asked to detect GPT-2 generated text in tweets from Twitter's publicly available disclosure of state linked information operations from 2019 and 2020, approximately 6% were predicted to be deepfakes.
Manual Review	However, a manual review of the tweets labeled deepfakes reveals that the majority of labeled tweets contained porn or spam. Also, a number of these tweets were very short, only a few words, which is where the model struggles to perform. Finally, when we ran the model on tweets from prior to GPT-2 release in 2019 (before GPT-2) we found roughly the same number labeled deepfakes. With these findings, we concluded that a majority of the set's deepfake labels were false positives.
Overall accuracy	We assess that because the rate of false positives is so high, the model's the rate of accuracy in an uncontrolled environment like Twitter, diminishes significantly, making the model virtually ineffective at detecting deepfake videos.

GAN Images	
Results of detection accuracy in controlled environment	A VGG16/CNN-based model, achieved 92% accuracy in a controlled environment.[64]

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

Results of detection accuracy in uncontrolled environments	We applied this model to images found in Twitter's publicly available Tweets from 2019 and 2020. This exercise required some preprocessing to detect (and subsequently crop out) the faces in the images. The detection model identified 6% of these images as deepfakes
Manual Review	Without a labeled dataset, we cannot say what the suspect accuracy of the pretrained detector model is. A manual review of a subset of the images labelled as deepfakes indicates that most are either drawn faces or faces that were extremely blurry, leading us to believe the vast majority of detections are false positives.
Overall accuracy	We assess that because the rate of false positives is so high, the model's the rate of accuracy in an uncontrolled environment like Twitter, diminishes significantly, making the model virtually ineffective at detecting deepfake videos.

Deepfake Video	
Results of detection accuracy in controlled environment	An XceptionNet based deepfake video detection model, trained on the FaceForensics++ dataset, takes frames from a video and labels them real or fake individually. It achieved 81% detection accuracy on low quality compressed deepfake videos.
Results of detection accuracy in uncontrolled environments	When tested on videos from Tweets from Twitter's publicly available 2019 and 2020 dataset, the XceptionNet based model performed poorly.
Manual Review	The detector model detected deepfake images in almost all videos.
Overall accuracy	We assess that this model is trained for a specific dataset (FaceForensics++) and cannot be accurately applied to uncontrolled environments such videos on Twitter.

(U) ENDNOTES

[1] Most Canadians have viewed some form of synthetic content on social media due to 1) the large amounts of synthetic content circulating on social media and 2) Canadians' high intake of social media content: Researchers at the Queensland University of Technology found that, on average, over 3.2. billion photos and 720,000 hours of video are created daily and available online. They note that plenty of this online content consists of synthetic media shared on social media. In 2018, 78% of Canadians used at least one social networking account and as of January 2021, the estimated number of Canadian users on social media platforms Facebook, Instagram, Twitter, TikTok, WeChat and Youtube totalled 67.1 million. See Sebastien Charlton and Kamille Leclair, Digital News Report: Canada 2019 Data Overview, Centre d'études des médias, Département d'information et de communication, Université Laval. February 2019; Schimmele et al. Study: Canadians' assessments of social media in their lives, Statistics Canada, 24 March 2021; T.J. Thompson et al., Visual Mis/disinformation in Journalism and Public Communications: Current Verification Practices, Challenges, and Future Opportunities, Journalism Practice, Research Gate, October 2020.

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

5 (U) Sebastien Charlton and Kamille Leclair, Digital News Report: Canada 2019 Data Overview, Centre d'études des médias, Département d'information et de communication, Université Laval. February 2019.
6 (U) FBI Private Industry Notification, Pin Number 210310-001, 10 March 2021.
7 (U) Nguyen et a. Deep Learning for Deepfakes Creation and Detection: A Survey, 26 April 2021, arXiv: 1909.11573v3.; Ian J. Goodfellow et al. Generative Adversarial Net. Département d'informatique et de recherche opérationnelle Université de Montréal, 10 June 2014.
8 (U) James Vincent, Microsoft invests $1 billion in OpenAI to pursue hayl grail of artificial intelligence, The Verge, 22 July 2019.;
(U) Amit Raja Naik, Kiss Me Transformer: A Journey of OpenAI, Analytics India Magazine, 2 November 2021.
9 (U) Wojciech Zaremba and Greg Brockman, OpenAI Codex, OpenAI Blog, 10 August 2021.
(U) Ramesh et al. DALL-E: Creating Images from Text, OpenAI Blog, 5 January 2021.
10 (U) Better Language Models and Their Implications, OpenAI Blog, 14 February 2019.
11 (U) Better Language Models and Their Implications, OpenAI Blog, 14 February 2019.
12 (U) Elizabeth Clark et al. All that's 'Human' is not Gold: Evaluating Human Evaluation of Generated Text, Paul G. Allen School of Computer Science & Engineering, University of Washington, Allen Institute for Artificial Intelligence, June 2021.
13 (U) Alex Newhouse, Jason Blazakis, Kris McGuffie, The industrialization of Terrorist Propaganda; Neural Language Models and the Threat of Fake Content Generation, Middlebury Institute of International Studies Center on Terrorism, Extremism and Counterterrorism, October 2019.
14 (U) When prompted with the sentence "the creators were worried about its ability to generate disinformation" GPT-2 come up with this response: "an extremely useful tool, but for us it could also be considered dangerous."
15 (U) OpenAI API, OpenAI Blog, 11 June 2020.
16 (TS//SI//REL TO

(U) Kris McGuffie, Alex Newhouse, The Radicalization Risks of GPT-3 and Advanced Neural Language Models, Middlebury Institute of International Studies Center on Terrorism, Extremism and Counterterrorism (CTEC), September 2020.

18 (U) In 2021, researchers from Singapore's Government Technology Agency demonstrated that, when given a choice to choose the legitimate email link, more people clicked on the link in the GPT-3 generated email than on the one in the email composed by a human. Lily Hay Newman, AI Wrote Better Phishing Emails Than Humans in a Recent Test, Wired, 8 August 2021. Accessed on 12 August 2021.

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

27 (U) Ian J. Goodfellow et al. Generative Adversarial Net. Département d'informatique et de recherche opérationnelle Université de Montréal, 10 June 2014.
28 (U) Ian J. Goodfellow et al. Generative Adversarial Net. Département d'informatique et de recherche opérationnelle Université de Montréal, 10 June 2014.

31 (U) Flora Carmichael, "How a fake network pushes pro-China propaganda" BBC News, 5 August 2021.
32 (U) Flora Carmichael, "How a fake network pushes pro-China propaganda" BBC News, 5 August 2021.

34 (U//OUO) FireEye, Network of Likely Inauthentic Pro-PRC Social Media Accounts Promotes Content in Support of New Hong Kong National Security Law and Portrays U.S. Black Lives Matter Protests as Chaotic, 2 July 2020.
(U//OUO) Mandiant Threat Intelligence, Pro-PRC Influence Activity Identified on Dozens of Additional Platforms, Websites, Forums, and in Three New Languages; Attempts to Physically Mobilize Protestors in the U.S. and Limited Authentic Engagement Observed, 11 June 2021.
(U//OUO) Mandiant Threat Intelligence, Threat Activity Report: Network of Pro-PRC Social Media Accounts Promotes Content Including That Critical of U.S. COVID-19 Vaccine Distribution 'Selfishness' Compared to International Distribution of Sinopharm Vaccine, 23 March 2021.
(U//OUO) Mandiant Threat Intelligence, Threat Activity Report: Network of Pro-PRC Social Media Accounts Promotes Content Critical of U.S., Including That Alleging U.S. 'Selfishness' Regarding India's COVID-19 Crisis, 12 May 2021.
35 (U) Yevgeny Kuklychev, Innovative information disorder tactics target Russia protests, First Draft News, February 2021.;
(U) Instagram deletes more than 500 accounts for scaremongering amid pro-Navalny rallies, Meduza, 4 March 2021.
36 (U) This included the former deputy director of former US President Donald Trump's domestic policy council, see Raphael Satter, Experts: Spy used AI-generated face to connect with targets, Associated Press, June 2019.

38 (U) Screen shots taken from original movie/tv show and deepfake versions of that movie/tv show. See Linda Carter Wonder Woman [deepfake], YouTube, 15 October 2020; Wonder Woman (2017) – No Man's Land Scene (6/10) – Movieclips, YouTube, 4 May 2018; Better Call Trump: Money Laundering 101 [DeepFake], YouTube, 18 September 2019; Saul Teaches Jesse Money Laundering – Kafkaesque – Breaking Bad, YouTube, 13 April 2021.
39 (U) Nguyen et a. Deep Learning for Deepfakes Creation and Detection: A Survey, 26 April 2021, arXiv: 1909.11573v3
40 (U) Adrian Tijie Xu, AI, Truth, and Society: Deepfakes at the front of the Technological Cold War. Gradient Crescent, Medium.
41 (U) Christian Vaccari and Andrew Chadwick, Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty and Trust in News, Social Media and Society, Volume 6, Issue 1, SAGE, February 2020.

44 (U) 2019 Update: Cyber Threats to Canada's Democratic Process, Canadian Center for Cyber Security, Communication Security Establishment, Cyber Threats to Canada's Democratic Process. 8 April 2019.
45 (U) Tim Hwang, Deepfakes: A Grounded Threat Assessment, Center for Security and Emerging Technology, July 2020.
46 (U) This phenomenon has been coined the "liar's dividend" meaning that deepfakes can make it easier for liars to deny the truth.
Tackling deepfakes in European policy, Panel for the Future of Science and Technology, European Parliamentary Research Service, July 2021.
47 (U) For example, in 2018 the Gabonese President Ali Bongo delivered a traditional New Year's address to the population after months of being absent from public life. President Bongo's unusual appearance in the video led many to speculate that the video was a deepfake and when members of Gabon's military attempted a coup against the government they stated that the video proved that something was wrong with the president. Later forensic analysis revealed that the video was not a deepfake and the government explained that the President's unusual appearance was due to the fact that he suffered a stroke prior to the video's filming, see Ajder et al, The State of Deepfakes: Landscapes, Threats and Impact, Deeptrace Labs, September 2019.

TS//SI//REL TO CAN, FVEY

CANADIAN CENTRE FOR CYBER SECURITY

50 (U//OUO) Deepfake Campaign Impersonating Russian Opposition Tricks Baltic Politicians and Media in an Attempt to Discredit West and Russian Activists, CrowdStrike, CSA – 210397, 28 April 2021.
51 (U//OUO) Deepfake Campaign Impersonating Russian Opposition Tricks Baltic Politicians and Media in an Attempt to Discredit West and Russian Activists, CrowdStrike, CSA – 210397, 28 April 2021.
52 (U//OUO) Vendor reporting has flagged that deepfake video and audio technologies consist of an emerging trend amongst criminals committing fraud and conducting business email compromise (BEC) attacks and other voice phishing (vishing) schemes, see
Deepfake Technology in Cyber Operations, Crowdstrike Intelligence Tipper, CSIT-21217, Crowdstrike Global Intelligence Team, 26 October 2021

54 (U) Tackling deepfakes in European policy, Panel for the Future of Science and Technology, European Parliamentary Research Service, July 2021.
55 (U) Huffington Post, A Powerful New Deepfake Tool Has Digitally Undressed Thousands of Women, 10 August 2021.
56 (U) The Business of Fraud: Deepfakes, Fraud's Next Frontier, Recorded Future, 29 April 2021; Shamani Joshi, They Follow You on Instagram, Then Use Your Face to Make Deepfake Porn in This Sex Extortion Scam, Vice News, 7 September 2021.
57 (U) Joel Khalili, Forget sextortion scams, we're more worried about deepfake ransomware, TechRadar, 2 June 2020; Shamani Joshi, They Follow You on Instagram, Then Use Your Face to Make Deepfake Porn in This Sex Extortion Scam, Vice News, 7 September 2021.
58 (U) Ajder et al. The State of Deepfakes: Landscapes, Threats and Impact, Deeptrace Labs, September 2019, P. 1

64 (U) Rossler et al., FaceForensics++: Learning to Detect Manipulated Facial Images, 25 January 2019, arXiv190108971.