2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

The Next Generation of Cyber-Enabled Information Warfare

Kim Hartmann
Conflict Studies Research Centre
kim.hartmann@conflictstudies.org.uk

Keir Giles
Conflict Studies Research Centre
keir.giles@conflictstudies.org.uk

Abstract: Malign influence campaigns leveraging cyber capabilities have caused significant political disruption in the United States and elsewhere; but the next generation of campaigns could be considerably more damaging as a result of the widespread use of machine learning.

Current methods for successfully waging these campaigns depend on labour-intensive human interaction with targets. The introduction of machine learning, and potentially artificial intelligence (AI), will vastly enhance capabilities for automating the reaching of mass audiences with tailored and plausible content. Consequently, they will render malicious actors even more powerful.

Tools for making use of machine learning in information operations are developing at an extraordinarily rapid pace, and are becoming rapidly more available and affordable for a much wider variety of users. Until early 2018 it was assumed that the utilisation of AI methods by cyber criminals was not to be expected soon, because those methods rely on vast datasets, correspondingly vast computational power, or both, and demanded highly specialised skills and knowledge. However, in 2019 these assumptions proved invalid, as datasets and computing power were democratised and freely available tools obviated the need for special skills. It is reasonable to assume that this process will continue, transforming the landscape of deception, disinformation and influence online.

This article assesses the state of AI-enhanced cyber and information operations in late 2019 and investigates whether this may represent the beginnings of substantial and dangerous trends over the next decade. Areas to be considered include: social media campaigns using deepfakes; deepfake-enabled CEO fraud; machine-generated political astroturfing; and computers responding to the emotional state of those interacting with them, enabling automated, artificial humanoid disinformation campaigns.

Keywords: *deepfake, disinformation, information warfare, malign influence, artificial intelligence, machine learning, emotional modelling*

1. INTRODUCTION

The year 2019 saw rapid developments in the use of machine-learning techniques to assist and amplify malign influence campaigns. Early in the year, "Katie Jones" was the first publicly identified instance of a deepfake face image used in a social media campaign.[1] By December, this technique had gone mainstream, with mass use in a campaign to influence US politics.[2] It is highly likely that this trend will continue in information operations, and as a result may transform the techniques, capabilities, reach and impact of information warfare.

Advances in artificial intelligence (AI) and the increasing availability of manipulation software usable by laymen have made the creation of convincing fake audio and video material relatively easy. The rapid spread of such material through social media and a lack of sufficient validation methods in cyberspace have resulted in the emergence of a potentially very powerful weapon for information operations. The speed of progress in this field is such that while deepfakes were not relevant for the 2016 US presidential election – at present the most prominent case study of cyber-enabled hostile interference in an election campaign – in 2020 they are widely regarded as a significant danger.

Until this point, malign influence and disinformation campaigns have primarily been operated and directed manually, or with the assistance of relatively crude and simple bots that are not able to interact convincingly with human targets or generate strategic long-term engagement. The design, production and dissemination of false material have been performed by human operators. But the trend of utilising AI methods to compose manipulated or fake material observed during 2019 indicates that it is possible to automate the processes needed to successfully operate disinformation

[1] Keir Giles, Kim Hartmann, and Munira Mustaffa, *The Role of Deepfakes in Malign Influence Campaigns*, (Riga: NATO STRATCOM COE, 2019), https://www.stratcomcoe.org/role-deepfakes-malign-influence-campaigns.

[2] Davey Alba, "Facebook Discovers Fakes that Show Evolution of Disinformation", *The New York Times*, 20 December 2019, https://www.nytimes.com/2019/12/20/business/facebook-ai-generated-profiles.html.

campaigns. In particular, this is because the level of sophistication of AI reached in a data processing and reasoning application context is different to AI in other fields. This type of AI may be considered as in between what are referred to as “strong” and “weak” AI. “Weak” AI is already available for generating specific output material or discrete tasks involved in disinformation operations, when the prerequisites and other inputs required to automate and generalise these tasks are already given. Currently these AI applications remain field-specific and hence cannot be considered as “strong” or true AI; however, with the appropriate supply of prerequisites and input data required to automate and generalise these tasks, their capabilities are much higher than the average “weak” AI already observed today.

While AI is still immature in many application scenarios, the technology has made significant steps in the specific areas of data analysis, classification, creation and manipulation, with a significant rise in the achievable output due to the availability of high-quality data and data processing routines (big data) as well as CPU power and memory capacities. While it is still difficult for AI systems to adapt to the real world, cyberspace – being an artificially generated domain constructed around pure data and communication – is their natural environment.

Most societies are still relatively accustomed to viewing audio and video recordings as indisputable evidence of reality. Images, video and audio recordings have played a major role in documenting our recent history and our trust in these recordings has shaped our perception of reality. Without modern media and our trust in them, our history is likely to have been different. An example is the release of the former US President Richard Nixon’s “smoking gun” tape, which eventually led to a change of power in the United States. Had this tape not existed, or had it not been trusted, history could have taken a completely different course.

In facing an era of artificially generated images, audio and video recordings, we are also confronted with the risk of real events being falsely claimed to be fake. As we currently do not have sufficient technologies to guarantee the authenticity of material being displayed, proving a falsely-claimed fake to be real may be even more challenging than the reverse. The effect of such false claims, especially in a political context, may be immense.

We have entered an era in which we depend heavily on audio and video materials as information resources while at the same time being confronted with the fact that this material can no longer be fully trusted. Although there have always been individuals who consider historic events such as the Holocaust, the moon landings or even the 9/11 terror attacks to be fictions despite multiple media evidence, current studies indicate that the number of individuals distrusting facts is rising rapidly due to the emergence of deepfake technology.[3] The erosion of trust in objective truth is accelerated by the ease with which apparently reliable representations of that truth can be fabricated; and augmented by the secondary effect of reduced trust in mainstream media, which neutralises their role in providing facts, informing the public and thus stabilising democratic processes. This plays directly into the hands of organised disinformation campaigns. A 2019 JRC Technical Report on the Case Study of the 2018 Italian General Election, published by the European Commission, indicated a correlation between distrust in media and a higher susceptibility to disinformation.[4]

Members of the US Congress have requested a formal report from the Director of National Intelligence on deepfakes and the threats they pose.[5] US senators Marco Rubio, member of the Senate Select Committee on Intelligence, and Mark Warner, Chairman of the Senate Select Committee on Intelligence, have urged social media companies to develop standards to tackle deepfakes, in light of foreign threats to the upcoming US elections. They note that: “If the public can no longer trust recorded events or images, it will have a corrosive impact on our democracy”.[6]

Meanwhile, by 2018 the US defence research agency DARPA had spent 68 million US dollars on a four-year programme developing digital forensics to identify deepfakes. However, there is a concern that the defending side in combating deepfakes will always be at a disadvantage. According to Hany Farid, a digital forensics expert at Dartmouth College: “The adversary will always win, you will always be able to create a compelling fake image, or video, but the ability to do that if we are successful on the forensics side is going to take more time, more effort, more skill and more risk”.[7]

[3] Karen Hao, “The biggest threat of deepfakes isn’t the deepfakes themselves - The mere idea of AI-synthesized media is already making people stop believing that real things are real”, *MIT Technology Review*, 10 October 2019, https://www.technologyreview.com/s/614526/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/; Simon Kuper, “The age of scepticism: from distrust to ‘deepfake’”, *Financial Times Magazine*, 18 October 2018. https://www.ft.com/content/2fc9c1fa-d1a2-11e8-a9f2-7574db66bcd5.

[4] Massimo Flore, Alexandra Balahur, Aldo Podavini, Marco Verile, “Understanding Citizens’ Vulnerabilities to Disinformation and Data-Driven Propaganda”, (Joint Research Centre (JRC) Technical Reports, European Commission, 2019), https://publications.jrc.ec.europa.eu/repository/bitstream/JRC116009/understanding_citizens_vulnerabilities_to_disinformation.pdf. On page 38 of the report it says: “They are designed to erode trust in mainstream media and institutions. Most of the content used to build these hostile narratives is not always objectively false. Much of it is not even classifiable as hate speech, but it is intended to reinforce tribalism, to polarize and divide, specifically designed to exploit social fractures, creating a distorted perception of reality by eroding the trust in media, institutions and eventually, democracy itself.”

[5] Donie O’Sullivan, “Lawmakers warn of ‘deepfake’ videos ahead of 2020 election”, *CNN Business*, 28 January 2019, https://edition.cnn.com/2019/01/28/tech/deepfake-lawmakers/index.html; Donie O’Sullivan, When seeing is no longer believing - Inside the Pentagon’s race against deepfake videos”, *CNN Business*, https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/.

[6] Marco Rubio website, “Rubio, Warner Express Concern Over Growing Threat Posed by Deepfakes”, 2 October 2019, https://www.rubio.senate.gov/public/index.cfm/2019/10/rubio-warner-express-concern-over-growing-threat-posed-by-deepfakes.

[7] Stephanie Kampf, Mark Kelley, “A new ‘arms race’: How the U.S. military is spending millions to fight fake images”, *CBC News*, 18 November 2018, https://www.cbc.ca/news/technology/fighting-fake-images-military-1.4905775.

In short, in the disinformation arms race the capabilities available to malign actors are developing and proliferating at an unprecedented rate, while states and others developing defensive or protective countermeasures are struggling to keep pace.[8] As seen in the field of cybersecurity in the past, the emergence of new threats such as AI-supported disinformation campaigns will not be avoidable.

The remaining sections of this paper explore the next generation of AI-enabled information warfare, and consider the acknowledged, but as yet vague and abstract, threat of weaponising AI for malign influence campaigns. Section 2 discusses methods and prerequisites for utilisation of AI in modern information warfare. Section 3 reviews the state of the art of AI capabilities for generating and processing different types of information material, including the ability to identify, respond to and generate emotional response in human-machine and human-computer interaction. Section 4 draws on the previous sections and past observations to look forward into the next decade of AI-enabled information warfare and possible countermeasures to it, and finally section 5 recommends steps that NATO member states should take to mitigate this new type of threat.

2. AI IN INFORMATION WARFARE

In order to understand the capabilities of AI-supported disinformation campaigns, it is necessary to understand what can be achieved by the technology used. The true power of AI in information warfare derives from several factors: societies' reliance on social media; dependence on cyberspace as a trustworthy information resource; unlimited access to and ability to spread information rapidly through cyberspace; and human difficulties in reliably distinguishing between fake and genuine media, as well as a lack of authentication or validation capabilities online.

Malign influence campaigns in 2019 and before have involved a wide range of material being manipulated through different techniques and targeting different human modalities.

A. Methods

While there are many methods within the field of machine learning that can be used for AI applications, generative adversarial neural networks (GANs) became prominent for deepfakes during 2019. GANs utilise neural networks to optimise their output. In simple terms, a GAN is a couple of neural networks playing a game against each other (most commonly a zero-sum-game in terms of Game Theory). In the case of deepfakes, one neural network aims at building a deepfake from a set of input data, while the other aims at correctly distinguishing the deepfake from the original data.

[8] Giles, Hartmann and Mustaffa, *The Role of Deepfakes*, 19–22.

Through this mechanism, the final output can be optimised with each "round" played. The method can be used both for the creation and alteration of media.

The artificial output produced becomes better over time and is also dependent on the required fidelity of the produced material. Typically, material of lower quality (image/video resolution or audio quality) is easier to fake, as there are fewer identifiable traits that must be learned. This has a direct effect on the amount of time needed for the training and hence on the time needed to produce a convincing deepfake.

From a technical perspective, there is a key difference between AI being used to create novel material and altering existing material. While the process involved varies slightly depending on the type of material being processed, the general concept remains similar. This allows an identification of the prerequisites needed, which is explored in the following subsection.

1) Creation

Currently, AI does not possess true creativity. Therefore, AI systems have problems generating unprecedented content, regardless of the type of output produced. However, what AI systems are particularly good at is learning correlations within data.

When producing novel content, AI systems tend to produce an average of the data used for training them. As an example: to produce a picture of an artificial woman, the AI will go through a database of images of women, extracting typical traits in those images in order to deliver an image containing the average of all identified traits. This is what most likely happened in the case of "Katie Jones". It also explains why she was identifiable as artificial through specific – yet minor – characteristics, such as blurred earrings of indefinable shape and colour. However, these artefacts of the AI processing can be avoided, either by manual post-processing of the generated output or by adjusting the AI accordingly.

Creating artificial content of a real, specific individual is slightly more complex and involves gathering training data on that particular individual. Publicly known individuals such as celebrities, politicians and major business leaders are therefore particularly at risk of being targets of AI-supported disinformation campaigns. Depending on the amount and quality of data available, creating artificial content may also involve application (and therewith learning) of general models, such as human-like movement patterns. This can then be used (to some extent) to compensate for a lack of sufficient data; however, it complicates the process and may be easier to identify as a fake.

2) Alteration

Compared to creation, alteration is somewhat more complex, as it involves more steps. In order, for example, to change a smile to a frown in a given picture, several steps are involved. First of all, the AI must understand which parts of a picture interact in order to be perceived as displaying a smile or a frown. These traits are universal to some extent, but may contain individual peculiarities. Hence, in order to be convincing, it is helpful to train the AI on the specific person whose image is to be altered. Having a model of how a smile (origin) and a frown (goal) look is the first step. The second step is to identify the areas that need to be altered. The third step is to perform the alteration and finally, the fourth step includes an adaptation to the overall image (such as light conditions, brightness and contrast). These steps are similar for video alterations.

Despite the fact that AI-supported alterations are somewhat more complex than generations, applications performing alterations already exist, as will be discussed further in section 3.

This kind of alteration should not be confused with simple editing, which continues to play an important role in disinformation. One prominent example from 2019 was a video of Ms. Nancy Pelosi, the US House Speaker and Democrat leader, which was altered to make her appear drunk and spread rapidly throughout social media.[9] This shows the potential of altered video material in misinformation campaigns generally. The case of Nancy Pelosi's altered video also showed some of the major concerns with social media. Despite the fact that the video gained 2 million views and had been shared 45.000 times within less than 36 hours,[10] Facebook confirmed that the video had been altered, but refused to take it down as "We don't have a policy that stipulates that the information you post on Facebook must be true."[11]

The Pelosi video was slowed down, making her speech appear slurred. Slowing down the replay rate of video and audio material is a very common task; most players have a function implemented for this purpose. Usually, the slower speed yields a notable change in the acoustics as well, resulting in a lower voice. In the case of Nancy Pelosi, the pitch had also been altered in order to compensate for this effect. While pitch alteration is not as common as change of the replay rate, it is still a task that requires little or no specific technical knowledge and is available on most audio and video processing applications. In a similar way, commonly available software for editing

[9] Joan Donovan, Britt Paris, "Beware the Cheapfakes", Slate.com, 12 June 2019, https://slate.com/technology/2019/06/drunk-pelosi-deepfakes-cheapfakes-artificial-intelligence-disinformation.html.

[10] Drew Harwell, "Faked Pelosi videos slowed to make her appear drunk, spread across social media", *The Washington Pos*t, 24 May 2019, https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/.

[11] Drew Harwell, "Pelosi says altered videos show Facebook leaders were 'willing enablers' of Russian election interference", *The Washington Post*, 29 May 2019, https://www.washingtonpost.com/technology/2019/05/29/pelosi-says-altered-videos-show-facebook-leaders-were-willing-enablers-russian-election-interference/.

still images, audio files and text will continue to play a key role in malign influence campaigns alongside more advanced technologies.

B. Prerequisites

The arrival of big data processing methods, advances in computational power and parallel and distributed computing means that machine learning is no longer an exquisite technology available only to actors with enormous resources.[12]

As the multitude of deepfakes that arose during 2019 showed, the technology to produce deepfakes has become widely available. Some applications, such as "Zao"[13] and "FaceApp",[14] are available for download, while others provide online service platforms to create deepfakes.[15] While it is unlikely that these applications will be directly used in a disinformation campaign, the technology is being offered as a business product and thus is at a level that allows it to be used by software developers to create according applications and may hence also be used to develop applications for malicious use-cases. In section 4 we examine further how such a weaponised AI for disinformation campaigns may look today and how it is most likely to be enhanced in the near future.

3. AI APPLICABILITY

One particularity of the AI methods utilised in disinformation campaigns is that they may be applied to basically any material available. The reason for this lies in their pure and abstract nature: as long as there are specific patterns identifiable in a set of data, an AI construct will be able to identify, learn and reproduce the correlations between them. In the field of disinformation, however, the most relevant media are text, audio and video, and consequently the following section will give a brief overview of the current state of manipulation and creation technologies for each of these forms of material, including the ability of AI to identify and display human emotion within this material. This in turn will enable a better understanding of their future potential for disinformation campaigns.

A. Text

While most attention has been devoted to the disinformation potential of manipulated video, audio and still images, artificially generated text has been feasible for over a decade. In 2008 SCIgen, a scientific paper generator programmed by MIT students,

[12] Samantha Cole, "This program makes it even easier to make deepfakes", vice.com, 19 August 2019, https://www.vice.com/en_us/article/kz4amx/fsgan-program-makes-it-even-easier-to-make-deepfakes; James Vincent, "AI deepfakes are now as simple as typing whatever you want your subject to say - A scarily simple way to create fake videos and misinformation", *The Verge Tech*, 10 June 2019, https://www.theverge.com/2019/6/10/18659432/deepfake-ai-fakes-tech-edit-video-by-typing-new-words.

[13] Zao app, https://www.zaoapp.net/.

[14] FaceApp, https://faceapp.com/app.

[15] Deepfakes web β, https://deepfakesweb.com/.

managed to generate a research paper that was accepted by a conference (Computer Science and Software Engineering, CSSE, 2008, co-funded by IEEE) practising peer-review for publication.[16] While the purpose of SCIgen was to "auto-generate submissions to conferences that you suspect might have very low submission standards",[17] it also shows the extent to which even longer plausible texts may be generated automatically. A more recent release on the topic of artificial intelligence being used to produce artificial texts is OpenAI's GPT-2.[18] The text generator has already been identified as having the potential to produce propaganda or misinformation by extremist groups.[19] The implications for malign influence campaigns are multiple, including reducing or removing the reliance on humans to generate interactions, and thus solving the problem of scalability. Astroturfing, the practice of fraudulently generating messages designed to give the impression of widespread support for an idea, becomes vastly easier when it is not necessary to manually craft each message.

B. Audio

In the context of disinformation campaigns, the utility of audio material used to impersonate another individual is self-evident. During 2019 audio deepfakes, utilising the same technology used to create fake videos, were generated to impersonate the voices of CEOs by fraudsters in cybercrime cases.[20] One particular case described in more detail by *The Wall Street Journal* led to a loss of USD 243,000 through a fraudulent bank transfer.[21] The case shows the potential of the technology as well as the vulnerability presented by our reliance on the auditory identification of individuals. If they demonstrate target-specific knowledge, phone callers are often accepted as legitimate without having gone through a sufficient identification process; this is even more the case if the conversation is not about financial transfers but political opinions or personal statements.

In addition, the availability of speech synthesisers and their ability to generate artificial voices that sound human are on the rise. These systems are even capable of adding emotional prosody to the speech produced.[22] Like the example of text above, the clear implication is that disinformation campaigns will no longer be constrained by

[16] The official "Herbert Schlangenman" blog, http://diehimmelistschoen.blogspot.com/.

[17] SCIgen homepage at PDOS research group of MIT CSAIL, https://pdos.csail.mit.edu/archive/scigen/.

[18] Irene Solaiman, Jack Clark, Miles Brundage, OpenAI Research Laboratory homepage and blog, "GPT-2: 1.5B Release", 5 November 2019, https://openai.com/blog/gpt-2-1-5b-release/.

[19] Liam Tung, "OpenAI's 'dangerous' AI text generator is out: People find GPT-2's words 'convincing' -The problem is the largest-ever GPT-2 model can also be fine-tuned for propaganda by extremist groups.", ZDNet.com, 6 November 2019, https://www.zdnet.com/article/openais-dangerous-ai-text-generator-is-out-people-find-gpt-2s-words-convincing/.

[20] Jesse Damiani, "A voice deepfake was used to scam a CEO out of $243,000", *Forbes*, 3 September 2019, https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.

[21] Catherine Stupp, "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case - Scams using artificial intelligence are a new challenge for companies", *The Wall Street Journal*, 30 August 2019, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

[22] Mark Schröder, "Emotional Speech Synthesis: A Review", Seventh European Conference on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark.

numbers; but in this case an additional challenge that will be overcome is linguistic ability. In early 2019, one of the authors was the subject of a crude attempt at social engineering to assist a cyber exploit, where spear phishing victims received a phone call from an individual claiming to be the author's personal assistant and urging them to click on the link they had just received. Several of the victims were made suspicious by the caller's thick Russian accent – but once AI-generated synthesised voice capabilities are available, this will no longer be a limiting factor.[23]

Less relevant to the explicit context of disinformation campaigns, but a convincing demonstration of the capacities of AI in audio processing, is AIVA (Artificial Intelligence Virtual Artist): an AI system composing emotional soundtrack music.[24]

C. Video

Deepfake video came to widespread attention during the course of 2019, whether created for entertainment purposes or to raise the public awareness of deepfakes and their potential. Examples included videos where items or individuals were added to an existing clip, as well as existing videos being altered and new ones created.

Video alteration has involved the use of mouth models to adapt lip and face movement to make the speaker appear convincingly to be delivering the fake speech on the audio track. While the technology behind this involves many disciplines ranging from video processing, movement and biodynamic modelling to audio processing, the orchestration of tools generated within these research fields has led to the creation of applications usable by laymen that are fully capable of producing convincing footage.

D. Images

Image manipulation applications have become almost omnipresent on social media platforms, ranging from applications used to enhance self-portraits to those that add, delete or alter content within a picture. Newer applications utilising AI enhance this capability by creating photorealistic images from simplistic drawings[25] or artificial images based on machine learning algorithms ("Katie Jones"). Images may also be used to create video footage (see section 3. C).

E. Emotional Response Patterns

At the time of writing, the authors are not aware of instances of alteration of emotional states being displayed in images and videos. Nevertheless, this capability should be easily within reach. The Human-Computer Interaction (HCI) research community has

[23] An overview of research projects in the field and their achievements can be viewed at http://emosamples.syntheticspeech.de/. The list is being maintained by Dr Felix Burkhardt, Director of Research at AudEERING GmbH (https://www.audeering.com/).

[24] AIVA -The Artificial Intelligence composing emotional soundtrack music, https://www.aiva.ai/, sample tracks of AIVA can be found on YouTube: https://www.youtube.com/watch?v=gzGkC_o9hXI.

[25] Nvidia AI Playground, Nvidia AI Research in Action, https://www.nvidia.com/en-us/research/ai-playground/.

devoted considerable effort to development both of systems capable of identifying, and virtual agents capable of displaying, human emotions. Videos simulating emotional reactions through facial movements are claimed to have been produced[26] from no more than a still image of a person and an audio clip.[27] The alteration of a video to include a simulated inappropriate emotional reaction could be a powerful tool to discredit public figures, especially as the changes made may be extremely subtle and hard to detect. A simple example could be a politician discussing a military operation that had claimed civilian victims, with his face altered to show indifference or even approval.

The HCI community has moved away from looking at what are known as the "Ekman basic emotions"[28] to concepts of more subtle emotional states and their transitions.[29] The research community has an excellent understanding of how emotions are being displayed and how to adapt systems to understand a specific users' hidden emotional cues.[30] However, with this knowledge, it is also able to reproduce footage displaying the subtle cues. Such alterations may even be difficult to identify for the individual being targeted, as many of these subtle emotional cues are a result of involuntary movements.[31]

4. THE NEXT DECADE

The weaponisation of AI for information warfare operations finds a natural home in cyberspace, an environment made up of pure digital data with no universal methods

[26] Konstantinos Vougioukas, Stavros Petridis and Maja Pantic, "Realistic Speech-Driven Facial Animation with GANs", *International Journal of Computer Vision* - Special Issue on Generating Realistic Visual Data of Human Behavior, Springer, Online 13 October 2019, https://link.springer.com/article/10.1007/s11263-019-01251-8.

[27] The video clips are available on YouTube: https://www.youtube.com/watch?v=NlNJKWPmmbk&feature=youtu.be.

[28] A set of emotions that are cross-culturally recognisable, which were defined by Paul Ekman and his colleagues in a 1992 cross-cultural study. The emotions identified were: anger, distrust, fear, happiness, sadness and surprise. These have become generally accepted within the HCI research community as the "basic emotions".

[29] Ingo Siegert, Kim Hartmann, Stefan Glüge and Andreas Wendemuth, "Modelling of Emotional Development within Human-Computer-Interaction", *Kognitive Systeme Journal* 2013, https://duepublico.uni-duisburg-essen.de/go/kognitivesysteme/2013/1/008; Kim Hartmann, Ingo Siegert, David Philippou-Hübner and Andreas Wendemuth, "Emotion detection in HCI: from speech features to emotion space." IFAC Proceedings 12th Symposium on Analysis, Design, and Evaluation of Human-Machine Systems, Volumes 46.15 (2013): 288–295, https://www.sciencedirect.com/journal/ifac-proceedings-volumes/vol/46/issue/15.

[30] Simon Peter van Rysewyk and Matthijs Pontier, *Machine Medical Ethics*, (Springer, 2014).

[31] Details on micro-expressions can be found through the Paul Ekman Group, a research group centred around Paul Ekman who also described the "Ekman Basic Emotions“ (see footnote 28) and has produced various publications on the topic, https://www.paulekman.com/resources/micro-expressions/; The Facial Action Coding System (FACS) is being used by HCI researchers worldwide to identify emotional user responses during the course of HCI; Paul Ekman, Erika L. Rosenberg, "What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)", (Oxford University Press USA, 1997).

for authentication and validation of data. This is likely to have a number of direct effects on the conduct or execution of information warfare.

A. Command and Control

While in 2019 the process of generating a deepfake required human intervention, over the next decade this will become a far more automated process.

During the 2010s a successful disinformation campaign needed humans at every stage. The concept had to be developed, and the material needed to be designed, drafted, generated and spread through social media platforms. Dissemination required the utilisation of social media profiles, which needed to be created in advance of the campaign, often manually. These profiles needed to be serviced by humans in order to build social networks, generate followers and establish credibility. Hence, disinformation campaigns involved human labour and indeed formed a whole disinformation industry in countries like the Philippines, India and Russia.[32]

Examining automation already in use on social media platforms today does suggest it is unlikely that this heavy reliance on a human workforce will continue. The individuals involved are most commonly low-budget service providers operating with limited resources. The engagement of such operators in disinformation campaigns has several drawbacks, the most prominent ones being that they may accidentally (or, as in the case of the Internet Research Agency in St. Petersburg, Russia, deliberately) disclose details of their activities[33] – but in general they are less effective at operating covertly and are less efficient.

Platforms such as Instagram are already known for the presence of bot activities. Ingramer[34] provides Instagram bot services that take over users' account(s) and allow fully automated, simulated human behaviour on the platform. Ingramer even ensures that it cannot be tracked by Instagram through geo-location metadata; it performs actions such as like/follow/unfollow, direct messages, scheduled post, hashtagging, location and username targeting.

Similar bots and processes exist on most social media platforms. They have become relatively easy to develop, since most services/application providers of social media platforms allow developers to interact with the platform through developer application programming interfaces (APIs). These allow software developers to interact with the platform's application/service through their own code/applications.

[32] Jonathan Corpus Ong, Jason Vincent A. Cabañes, *Politics and profit in the fake news factory – Four work models of political trolling in the Philippines*, (Riga: NATO STRATCOM COE, 2019), https://www.stratcomcoe.org/four-work-models-political-trolling-philippines.

[33] EUvsDisinfo.eu, "Confessions of a pro-Kremlin troll", 26 April 2017, https://euvsdisinfo.eu/confessions-of-a-pro-kremlin-troll/.

[34] Ingramer-Bots homepage, https://ingramer.com/.

Applications that address several APIs of different social media platforms are capable of controlling multiple accounts on multiple platforms. Such applications already exist and are available online. They are generally referred to as "social media management apps", and include examples such as Agorapulse,[35] Sprout Social[36] or Hootsuite.[37] The latter has a list of apps available that allow a connection of bots to Hootsuite Inbox.[38]

Since control panels for automated postings on social media are a mature, widely and cheaply available and broadly accepted technology, development of "command and control" panels for disinformation operations in hybrid warfare should be expected. Combining these with parallel developments in machine learning makes it likely that they will control AI agents (intelligent bots) capable of generating artificial content (semi-) automatically. The benefits are evident: a potentially unlimited number of accounts on multiple social media platforms that can be orchestrated by one individual, through one single application, spreading content generated by artificial intelligence pursuing a single and coordinated strategic goal.

B. Scalability

The technology used to produce deepfakes and other manipulated material is, at its core, nothing other than software. One goal for the weaponisation of AI for information warfare purposes in social media spaces is to automatically produce content that is coherent with the overall strategy of a disinformation campaign, but uses different means to display, share, and interact with the content produced. Due to the way social media works, this will heighten the trustworthiness of the content produced and ensure wide dissemination of the material.

As the scalability of software has been a major concern to the software engineering industry over the past years, especially with the shift towards "as a service" architectures, many concepts have been developed to allow an easy scaling of necessary software components. One of these concepts is microservice architectures, where each component of the software is a separate entity capable of operating on its own. This concept works very well with that of software agents. These entities (microservices) interact and respond to higher demands by automatically deploying several copies (instances) of themselves automatically through a so-called CI/CD (continuous integration/continuous deployment) pipeline. The CI/CD pipeline is part of "DevOps" (development operations) and the use of microservices with automated deployment is already industry standard for software engineers working on cloud architectures and other services needing to respond to changing demands.

[35] Agorapulse: Social Media Management Software for Agencies and Teams, https://www.agorapulse.com/.

[36] Sprout Social: Social Management Solution, https://sproutsocial.com/.

[37] Hootsuite Social Media Tool – Schedule your Tweets, https://hootsuite.com/.

[38] Hootsuite Apps – Bots – Apps that allow you to connect bots to Hootsuite Inbox, https://apps.hootsuite.com/categories/bots.

When designing a "command and control" panel as described above, it is reasonable to use a software engineering pattern that allows scalability. This will yield a more robust platform capable of producing high through- and output, where one panel is able to control hundreds of apparently independent accounts managed by software agents. This will ensure that performance limitations are negligible and allow a spontaneous adaptation to changing demands. The remaining limiting factor will be the control mechanisms installed by social media platforms, which are currently known to be insufficient.[39]

C. Automation

AI-assisted automation is very likely to be a major feature of the next decade of information warfare. This could apply in two distinct fields: automatically releasing disinformation following a coordinated overall strategy, and the automation of generating the disinformation. The latter task depends on acquisition of the data needed for the AI methods and their capability to generate content, preferably following a specific strategy (such as propaganda involving racism against a specific ethnicity). Automated release of already-available disinformation is easier to achieve, as it only requires scheduled access to the platforms targeted. From a technical perspective, this does not necessarily involve any artificial intelligence, although AI may be beneficial in order to create a more realistic illusion of human behaviour.

Automation could in the future also be used to generate instant responses to events. An intelligent information warfare campaign should be able to identify the rising interest in a relevant topic (such as the popularity of a specific individual or action) and generate a coordinated automatic response to leverage the interest. Response patterns could include producing counterarguments, fake news, "trolling" or cheap propaganda. In this context, the already existing abilities of AI systems to identify emotional states being displayed, to produce emotionally coloured responses, and to foresee their effects on humans will become of particular value. At present, all of these require a high number of user accounts sharing or promoting the produced material, which provides an obvious role for automation by more sophisticated means than the bots currently in use.

While the process of generating deepfakes is currently still being initiated manually, it should be expected that this too may soon be automated. However, producing still images to generate a profile picture of an artificial individual such as "Katie Jones" will still be far simpler than automatically generating a convincing deepfake video of an existing individual to deliver an automatically generated speech. It is likely that this type of activity will still involve a human workforce for the time being,

[39] Sebastian Bay and Rolf Fredheim, *Falling behind: How social media companies are failing to combat inauthentic behaviour online*, (Riga: NATO STRATCOM COE, 2019), https://www.stratcomcoe.org/how-social-media-companies-are-failing-combat-inauthentic-behaviour-online.

until AI systems are capable of acting according to an abstract goal such as the one a disinformation campaign may have.

As described in section 3. A, the production of shorter texts with the aid of AI when given a set of keywords is already reality. Having bots active on social media platforms that post these artificially generated texts is not a challenge. Even today, social media users such as influencers manage their account(s) through applications that allow them to schedule pre-defined posts or to generate posts out of a set of texts, hashtags and pictures. It is likely that similar techniques, enhanced through machine learning, will be deployed in information warfare in the near future, augmenting troll factories and botnets.

D. Countermeasures

It is important to understand that the processes described in this paper do not depend on future or emerging technologies. Each of the capabilities required is already at an advanced stage, and the respective research fields have developed these technologies for specific and multi-modal systems over the past years and, in some cases, decades. The techniques are ready for use in legitimate civilian applications, and in many instances are already known to be being weaponised by malicious actors. This is a matter for real and urgent alarm, since AI-supported disinformation campaigns have the potential to impose the largest threat to democracy and society seen so far, targeting not only public opinion but the nature of belief and trust, which constitute pillars of democratic societies.

In common with other new technologies, it is very unlikely that weaponisation can be prevented. Instead, methods to authenticate and distinguish original from manipulated material on a mass scale and in real time are urgently required. The particular problem with identifying manipulated material lies within the methods used to generate this material: GANs, as described above, if sufficiently well-trained, will yield an outcome that is difficult to distinguish from genuine media – not just for the human observer, but also for machines.

An alternative approach could be certification of genuine material. In the same way as the internet as a whole was designed to be insecure, meaning that secure applications and processes needed to be developed separately, so in an information space that is generally untrusted, additional measures could be necessary to stimulate trust. Possible technologies for doing so could include digital watermarks (requiring the involvement of manufacturers to include the technology in recording devices), or software signature processes, as known in e-mail communication. In both cases, however, this kind of approach would only be of use for a small subset of the total amount of information in circulation. The disinformation industry relies heavily on propaganda being spread through social media. The material being spread does not necessarily have to appear official, as long as it is convincingly real, or at least plausible, and provides an explanation of how or why someone got access to the record. At the same time, the widespread consumption of this type of material has contributed to the public becoming accustomed to low-quality material originating from doubtful sources and claiming to show the real truth to a story. This eases the task of malicious actors intending to spread disinformation.

A third approach that requires further investigation is that of using distributed knowledge to validate material being circulated. The idea is to use the knowledge of several individuals, sensors and general information, combined to reason the validity of the material being displayed. This combined knowledge could include verification by known witnesses, physical phenomenon validity checks (e.g. light effects or interactions between the environment and objects in videos), surveillance monitoring data and background information checks (such as validation of the caller ID in telephone calls through the service provider or specific knowledge of the location being filmed).[40] Notable results may be derived from research into swarm intelligence, a subfield of artificial intelligence.

5. OUTLOOK

Information warfare lies at the intersection of several well-established trends that will combine to pose severe challenges to nations and societies in the short and medium term. These are:

- The continuing progress of hyperconnectivity, reducing the perceptibility of dividing lines between online and real life;[41]
- Reduced restraint by actors hostile to liberal democracies, as they are emboldened by the apparent lack of deterrent measures available to their targets;
- Further erosion of trust, and of the notion of independent and verifiable truth;[42]
- Finally, as detailed in this paper, the rapid and accelerating pace of change in technologies that facilitate or enable malign influence campaigns.

[40] Jack Corrigan, "DARPA Is Taking On the Deepfake Problem", NextGov.com, 6 August 2019, https://www.nextgov.com/emerging-tech/2019/08/darpa-taking-deepfake-problem/158980/, "A comprehensive suite of semantic inconsistency detectors would dramatically increase the burden on media falsifiers, requiring the creators of falsified media to get every semantic detail correct, while defenders only need to find one, or a very few, inconsistencies,"; Derek B. Johnson, "The semantics of disinformation", Defensesystems.com, 26 August 2019, https://defensesystems.com/articles/2019/08/26/darpa-disinformation-semantics-johnson.aspx.

[41] Kim Hartmann and Keir Giles, "Shifting the core: How emergent technology transforms information security challenges", *Datenschutz und Datensicherheit (DuD)*, Springer Journal, 14 June 2017, https://link.springer.com/article/10.1007/s11623-017-0807-y.

[42] Giles, Hartmann and Mustaffa, *The Role of Deepfakes*.

Of these parallel but interdependent phenomena, perhaps the most straightforward to prepare for is the impact of technologies. Unlike the other trends, this is both relatively predictable and has a set of clearly identifiable countermeasures.

These could include:

- Exploring methods of technical authentication of digital material;[43]
- Content provenance through digital signatures;[44]
- Considering further applications of digital signatures;[45]
- Continuing efforts to restore trust in independent media and journalism;
- Inducing social media platforms to enhance the detection of fakes and to install means to allow users to evaluate the reliability of content;
- Ensuring the availability of national or supranational authorities to which civilians can report instances of malign influence campaigns;
- Following the example of Singapore,[46] forcing social media platforms to mark fake or false content (including any repost or shared post of the initial material).

Each of the new technologies detailed in this paper will have an impact on information warfare; but it need not be a transformative one. As with other previous technological developments, delivery of disinformation may be effected in a different manner, but the fundamental nature of deception remains unchanged. As such, the basic ingredients of countering it follow the same pattern as in previous decades and indeed centuries. This is because while disinformation techniques and technologies change, one factor that remains constant is the human susceptibilities they exploit.

It follows that alongside the technical recommendations above, individual states should undertake clear and honest assessments of their own publics' susceptibility to malign influence campaigns. Metrics to quantify and understand this susceptibility are as urgently needed as metrics to assess the success or failure of disinformation campaigns, and are essential both to preparing countermeasures and to fostering societal awareness of the threat.

A wide range of factors determine how susceptible a community is to disinformation: access to and engagement in social media, media uptake and trustworthiness, age,

[43] Antonio García Martínez, "The blockchain solution to our deepfake problems", *Wired Magazine*, 26 March 2018, https://www.wired.com/story/the-blockchain-solution-to-our-deepfake-problems/.

[44] National Academics of Sciences, Engineering, and Medicine, Chapter 6 "Deepfakes" in *Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop*, (The National Academies Press, Washington D.C. 2019), https://doi.org/10.17226/25488.

[45] Kalev Leetaru , "Why digital signatures won't prevent deep fakes but will help repressive governments", *Forbes*, 9 September 2018, https://www.forbes.com/sites/kalevleetaru/2018/09/09/why-digital-signatures-wont-prevent-deep-fakes-but-will-help-repressive-governments/.

[46] Singapore Legal Advice, "Singapore Fake News Laws: Guide to POFMA (Protection from Online Falsehoods and Manipulation Act)", 2 January 2020, https://singaporelegaladvice.com/law-articles/singapore-fake-news-protection-online-falsehoods-manipulation/.

technical education level, trust and understanding of democratic values, as well as trust in national leaders. The latter point places an obligation on leadership figures in Western liberal democracies to understand their own contribution to societal cohesion and common defence. Trust in leaders and institutions is a foundation stone of democratic systems; and an erosion of this trust through flagrant disregard for honesty and probity while in power creates a power vacuum which can and will be exploited by adversaries.

In keeping with all of this, a recommendation that remains common to all counter-disinformation efforts is raising public awareness: of the threat, of its methods, and of the indicators and warnings that an individual or group is being subjected to a malign influence campaign – critically, regardless of whether this campaign is mounted by foreign adversaries or domestic political actors manipulating society for their own ends. To this long-standing recommendation should now be added audience education on the nature and capabilities of the next generation of AI-enabled disinformation technologies. Proper preparation and investment in threat literacy among target audiences, started now, will have a substantial impact in mitigating the potential dangers of information warfare in the next, even more complicated decade.