

The weaponisation of synthetic media: what threat does this pose to national security?

Matteo E. Bonfanti | Senior Researcher in the Cyberdefense Project with the Risk and Resilience Team at the Center for Security Studies (CSS) at ETH Center for Security Studies, Zurich | @Teobonf05 

Theme

The weaponisation of hyper-realistic synthetic video, audio, images, or texts –generally known as of synthetic media– may affect national security. How and to what extent?

Summary

The article deals with the national security implications generated by the weaponisation of AI-manipulated digital content. In particular, it outlines the threats and risks associated with the employment of hyper-realistic synthetic video, audio, images or texts –generally known as ‘synthetic media’ or ‘deepfakes’– to compromise (influence) targeted decision-making processes which have national security relevance. It argues that synthetic media would most likely be employed within information and influence operations targeting the public opinion or predefined social groups. Other potential national security relevant targets (specific individuals or organisations that have responsibility for national security) should be adequately equipped to deal with the threat in question (and therefore have appropriate procedures, technologies and organisational settings to deal with the threat).

Analysis

Introduction

The malicious manipulation of information is not a new phenomenon.¹ It has ancient origins, although it has evolved over time.² It has developed in parallel with the gradual transformation of society, driven by scientific and technological innovation. In concrete terms, the malicious manipulation of information has gradually changed in terms of the methods, techniques and tools employed, although the goals pursued by the instigators are essentially the same: to compromise the individual or collective interest to the correct and complete representation of reality.³

¹ The expression here refers to the intentional production or reproduction of information that is false, altered, fragmented or decontextualised for misleading purposes or, in any case, to cause harm or damage to third parties.

² J.-B. Jeangène Vilmer, A. Escorcía, M. Guillaume, J. Herrera, “Information Manipulation: A Challenge for Our Democracies”, Report by the Policy Planning Staff (CAPS) of the Ministry for Europe and Foreign Affairs and the Institute for Strategic Research (IRSEM) of the Ministry for the Armed Forces, Paris, August 2018.

³ See M.E. Bonfanti, “An Intelligence-Based Approach to Countering Social Media Influence Operations”, in *Romanian Intelligence Studies Review*, No. 19-20, 2018, pp. 47-67.

Nowadays, the so-called information and communication society faces a further iteration of the phenomenon. Due to the increasing availability of the Internet as well as of technological devices enabling their users to be constantly connected, manipulative actions have invaded the cyberspace. These actions count on digital content that is delivered in and through virtual platforms, including social networks. The content is generated or altered using approaches, techniques and instruments that involve varying levels of complexity and innovation. These include emerging machine/deep-learning solutions that, according to some experts, are quickly pushing the manipulation of digital information to a level of sophistication and risk that would have been inconceivable until few years ago.

In this regard, quite concerning are those solutions based on deep-learning techniques, those that make use of so-called Generative Adversarial Networks (GANs). GANs are two systems of competing artificial neural networks –the generator and the discriminator– that allow the creation of hyper-realistic video, audio, images or text.⁴ Such digital content can only be identified as inauthentic through careful and painstaking examination, which is enabled by using conventional and –increasingly– non-conventional forensic techniques (technological tools that in turn incorporate artificial neural networks). Once generated, the content –known as ‘synthetic media’ or ‘deepfakes’– can be abused.

As widely documented by the media, abuse mostly consists in the unauthorised creation and sharing of pornographic images and videos, which impair the dignity, honour, and reputation of the victims. Abuse often entails cyberbullying, stalking or defamation via on-line media.⁵ Less frequent (although on the rise) are cases involving the creation and deployment of synthetic media for blackmailing, extortion and fraud that are directed towards both individuals and organisations. With regard to frauds against organisations, they are perpetrated by resorting to the well-known technique of the ‘CEO scam’ or ‘CEO fraud’, which is upgraded thank to the use of a deep fake audio.⁶ Although hypothetical, synthetic digital content could also serve as (inauthentic) documentary evidence to commit insurance or other types of fraud, or even to taint court cases. Less hypothetical it is their employment to sway public opinion. Indeed, synthetic media are suitable to political propaganda or disinformation. There are already few examples of this type of use. One may remember the altered video produced by the Belgian Socialist Party, which showed the American President Donald Trump inviting the Belgian Government to withdraw from the Paris climate agreement.⁷ Or the fake video ridiculing the Italian senator Matteo Renzi which was also broadcasted by a national television channel.⁸ Or,

⁴ A. Collins, “Forged Authenticity: Governing Deepfake Risks”, EPFL Policy Brief, International Risk Governance Center, 2019.

⁵ This content is often created through ‘face-swapping’, a technique that involves replacing the face of an actor with the face of an unwitting victim.

⁶ In 2019, there were cases about personnel from certain credit institutions who were misled by audio – generated using deep-learning techniques – that reproduced the voice of legitimate CEOs. They were tricked and transferred funds to recipients who were not authorised to receive them. C. 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/VIII/2019.

⁷ H. Von Der Burchard, “Belgian socialist party circulates ‘deep fake’ Donald Trump video”, *Politico*, 21/VI/2018.

⁸ S. Cosimi, “Il deepfake di Renzi è un gioco molto pericoloso”, *Wired*, September 25, 2019.

again, the altered video representing Jeremy Corbyn and Boris Johnson who mutually supported each other's run for Prime Minister.⁹

Considering the above, one may wonder whether individuals, organisations and, more generally, public opinion are equipped to deal with the possible misuses of synthetic media. The question is complex and, in abstract terms, relates to the current degree of society's resilience to manipulative operations deploying hyper-realistic but fake representations of reality. As some commentators have observed, the perception that nothing can now be believed because of the possibility of creating 'authentically false' or 'falsely authentic' digital content would have a destabilising impact on civil society as a whole; an impact greater than the effect that would result from an actual abuse of these media. However, this is difficult to establish in practice. Among other things, the absence of, or difficulty in, identifying clear and, above all, valid parameters for measuring the above impact limits the possibility of assessing and quantifying its magnitude. That said, and despite the abovementioned limitation, it is anyhow interesting to ask how the communities can equip themselves to face the threats and risks associated with deepfakes.

The abuse of synthetic media: a growing concern for national security

The abuse of synthetic media can be (loosely) represented as per the following. In terms of the type of victims, abuse may involve individuals, organisations, and the civil society. It may consist in actions that are: illegal, if they unambiguously breach binding rules (for example, regulations on defamation or stalking); illegitimate, if they are not illegal but nonetheless incompatible with protected interests and values; and national security-related, when the protected interests relate to elements and activities that, if compromised, could jeopardise State's security, irrespective of the type of victims involved.

Many scholars, practitioners as well as representatives of various governmental bodies have recently discussed the national security-related implications of deep fakes. In this regard, it is worth mentioning the public hearing held by the US House Permanent Select Committee on Intelligence in June 2019.¹⁰ One may also recall the statement of the Director of the Pentagon's Joint Artificial Intelligence Center, according to whom 'Deepfakes are a national security issue'. This declaration followed on the comments already made by the Director of National Intelligence, Daniel R. Coats, who noted that 'Adversaries and strategic competitors probably will attempt to use deep fakes or similar machine-learning technologies to create convincing - but false - image, audio, and video files to augment influence campaigns directed against the United States and our allies and partners'.¹¹ Various commentators, experts and analysts have also seconded these conclusions. However, most of the (especially early-date) analytical contributions to the

⁹ BBC News Channel, 12/XI/2019, <https://www.bbc.com/news/av/technology-50381728/the-fake-video-where-johnson-and-corbyn-endorse-each-other>.

¹⁰ US House of Representatives, Hearing on National Security Challenges of Artificial Intelligence, Manipulated Media and Deepfakes, 13/VI/2019, <https://intelligence.house.gov/calendar/eventsingle.aspx?EventID=653>.

¹¹ Worldwide Threat Assessment of the US Intelligence Community, Statement for the Record, 29/I/2019, p. 7, <https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf>.

topic do not seem fully convincing. They do not show and clarify how and to what extent the phenomenon is relevant for national security.

Synthetic media: 'democratisation' vs 'weaponisation'

There are two aspects that are usually considered relevant in assessing the nature and extent of the threat and risks synthetic media might pose to national security (elements that, among other things, are also valid when national security is not at stake). The first relates to the rapid development and availability of machine- and deep-learning technologies and techniques (used to generate synthetic media) among communities of researchers and users distributed in different geographical areas and institutional sectors. The second concerns the possibility of misusing those technologies and techniques, to compromise essential national security interests. The first aspect refers to the 'democratisation' of scientific knowledge and technological/technical capabilities to generate synthetic media, while the second relates to their 'instrumentalisation' or 'weaponisation'.

'Democratisation' concerns the gradual and increasing availability –often facilitated by commercial applications or services– to a range of different actors of algorithms and models, advanced computational capabilities, and training data for neural networks. These are essentially the fundamental components required to generate synthetic media. In other words, democratisation refers to the process of scientific-technical knowledge transfer, which goes hand-in hand with the lowering of the entry barriers for acquiring or using the technologies to generate deepfakes.¹² However, although concerning, the process of democratisation and its implications should be taken with a grain of salt. Indeed, while it is true that the techniques and technologies used to generate deepfakes are becoming a sort of 'commodity', it seems plausible that the creation of high-quality/sophisticated false information nonetheless requires technical and other resources which are currently available to a very limited number of actors.¹³ Of course, this does not prevent these resources will at a certain point become accessible to a wider range of individuals or organisations. On the other hand, depending on the situation, the high degree of sophistication and quality might not be required to pursue a desirable goal.

Turning to the 'weaponisation' (whether current or potential) of synthetic media, this is an aspect which is frequently given less attention or is addressed in a more fragmentary manner by experts and practitioners. Most of the analytical contributions delivered so far have generally focussed more on the availability of technical resources for generating deepfakes than on the possibility/risk that such digital content would effectively be weaponised. Although commonplace, such an availability describes an aspect that should be considered in assessing the deepfakes-associated threat. However, the will and interest of certain actors to make concrete use of those capabilities to compromise national interests and security are no less important. While it is true that deepfakes can

¹² Deeptrace Labs., "The State of Deepfakes Landscape, Threats, and Impact", 4/IX/2019, p 4.

¹³ . Leslie, N. Hoad and B. Spraggon, "How hard is it to make a believable deepfake?", *ABC News*, 27/IX/2018, <https://www.abc.net.au/news/2018-09-28/fake-news-how-hard-is-it-to-make-a-deepfake-video/10313906>.

be employed for malicious purposes, it is equally true that such potential does not inevitably and automatically results in their actual use. At first sight, this clarification might seem trivial. And yet it is not, especially if one considers how the debate on deepfakes originally developed. Focusing on weaponisation is paramount to understand and better assess the scale of the threat and the risks associated with synthetic media.

It means that risks and threat assessments should be premised upon two framework questions: 'Can a given actor generate or have available deepfakes which could compromise national security, irrespective of the manipulated digital content quality and sophistication?'; 'Why and how could that actor weaponised deepfakes?'. The approach proposed here seems valid, also considering the above described process of democratisation that will make the capabilities required to generate synthetic media available to many actors. If many actors will have access to these capabilities, it will then become particularly important to provide a response to the second question above. The validity of the approach seems also supported by the practice. Indeed, taking into account the fact that the technology required to generate this type of content has been available since at least 2017, it is interesting to note that –to date, as far as we know– there have been no episodes of misuse intended to undermine the national security of a State.¹⁴

This is something that could have happened, for example, in the case of distribution of deepfakes relating to political leaders in the act of committing reprehensible actions (being corrupt or consuming illegal narcotics) or depicting representatives of security or armed forces committing abuses. Conversely, what has happened so far is the dissemination of audio - (such as the audio tape which documented an alleged episode of international corruption run by a representative of the Italian Lega Nord during a meeting at the Metropol Hotel in Moscow) or video (the video of the Austrian Vice-Chancellor and leader of the nationalist party FPÖ) –which are authentic until proven otherwise and that raised questions about the national security of the States concerned.¹⁵ This suggests that, because the deepfake phenomenon is becoming part of the public domain, its potential for damage and the surprise effect have probably been partially 'defused'. It means that some actors might have decided that it is in fact counterproductive to resort to synthetic media given that there are less sophisticated, "safer" and effective ways to manipulate information and obtain the desired effects. Reference is made to 'dumb', 'cheap' or 'shallow' fakes such as the videos of Nancy Pelosi or Jim Acosta, which were altered without the use of deep-learning techniques and nonetheless provoked much public discussion (although they did not raise major questions in terms of national security). Lastly, one should not overlook a fundamental principle about information warfare: it is advisable to avoid using information that is completely false and altered, unless it is strictly necessary and functional to achieve a given objective. Indeed, if recognised as fake –and this could also be the case for deepfakes–, both the sources responsible for disseminating the information and the

¹⁴ K. Giles, K. Hartmann, and M. Mustaffa, "The Role of Deepfakes in Malign Influence Campaigns", NATO Strategic Communications Centre of Excellence, 2019, pp. 9-11.

¹⁵ F. Sarzanini, "Salvini, Savoini e quella cena a Mosca alla vigilia del Metropol", *Corriere della Sera*, 19/VII/2019. T. Mastrobuoni "Austria, il video che imbarazza il Cancelliere: il vice fa affari con i russi", *La Repubblica*, 17/V/2019.

information itself lose credibility. The impact of debunking could therefore be greater than the impact of the fake content.

In the light of the above, it is nevertheless interesting to examine how synthetic media could be used to compromise national security. The next sections will tackle the issue. They will address the hypothetical abuse of deepfakes to compromise national security irrespective of the possibility such abuse might prove effective. In general, the effectiveness of possible weaponisation will largely depend on the level of preparation/resilience of the potential victims.

Weaponisation of synthetic media and national security

The weaponisation of synthetic media could affect State's national interests and, in particular, security. Given the breadth of the notion of national security, the weaponisation may in principle address varied security-related aspects/domains, including civil, military, domestic and foreign security. It could also involve different dimensions or segments of security, i.e. political, economic, territorial, and cyber-security.

In abstract terms, the abuse of synthetic media represents a threat to national security to the extent it is able to compromise –directly or indirectly– one or more essential interests of a State (sovereignty, stability, integrity, peace, economic prosperity), essential values (democracy, rule of law, liberty) or the independence and operation of relevant institutions. In particular, such use would become significant if it were intended to disrupt those decision-making processes that, if compromised, could result in a course of action that is harmful for the national community, its existence and cohesion.

The most plausible (although not yet occurred) employment of deepfakes would be within information and influence operations (in peacetime and during conflict, in ordinary or emergency situations).¹⁶ These are operations that use information to influence decision-making processes. They can target individuals or selected organisations (individual-/organisation-oriented), or public opinion and specific groups within the civil society (social-/community-oriented). Both types of operations can be connected and, in some cases, may overlap in the achievement of their operational goals. It means that although the direct, material targets of social-/community-oriented influence operations are public opinion or predefined social groups, specific decision-makers might be secondary, indirect targets. From this standpoint, influencing and mobilising civil society and public opinion through the use of influence operations incorporating deepfakes could contribute to exerting pressure and shaping the decision-making process of individuals or selected organisations. In terms of the actors that can foster, promote, and conduct operations of this type, the most dangerous are the so-called Advanced Persistent Manipulators (APMs). These are civilian or military security bodies, activist groups, subversive or extremist groups, political/party organisations, economic establishments or other private

¹⁶ Information and influence operations make use of tools and techniques intended to shape the ideas, opinions, attitudes, emotions and perceptions of a target in respect of a given situation/question, so as to induce that target to adopt a certain type of behaviour (e.g. undertake or refrain from undertaking certain actions).

organisations able to create a threat that is significant because it is lasting, sophisticated, widespread, generally latent, and adaptable/changeable.¹⁷

However, the possible weaponisation of deepfakes does not only involve the actions described above. Deepfakes could also be deployed within operations that are primarily intended to collect sensitive information. One should remember the case of Katie Jones, a fake Linked-in user whose profile image proved to be synthetically generated.¹⁸ Regardless of the real intentions of the creators/managers of the profile, the case shows how synthetic media can be instrumentalised to obtain, through deception, information that is relevant for achieving certain goals. Furthermore, synthetic media could be employed to circumvent authentication, recognition and control systems used to access protected areas. These systems include those integrating biometric recognition applications (voice or image) like scanners to verify the authenticity of identity documents.

Information and influence integrating synthetic media

The weaponisation of deepfakes could occur as part of information and influence operations targeted at individuals or organisations (government authorities or those involved in political representation, civilian or military bodies, providers of critical infrastructures or essential services) which, because of their roles, responsibilities, functions or specific areas of competence, make and implement decisions that are significant for national security purposes. In this case, the altered digital content would become the tool for influencing the decision-making process through deception, extortion, or coercion.

When used for deceptive purposes, synthetic media could support third party's impersonation. In operational terms, deception is implemented using methods like those characterising the 'CEO scam'. The use of synthetic media is intended to induce the victim to undertake (or refrain from undertaking) certain actions, such as transferring information or starting a given process.¹⁹ Synthetic media could also be deceptively used as false documentary evidence, once again to induce decision-makers to undertake or refrain from undertaking certain actions. For example, this could involve the manipulation

¹⁷ M.E. Bonfanti (2020), Spazio Cibernetico e Operazioni di Influenza: Profili Evolutivi della Minaccia e Attività Informative di Contrasto, in U. Gori, Information Warfare, forthcoming.

¹⁸ The LinkedIn profile of Katie Jones, a self-styled researcher at a US think tank, appeared on the social media platform in March 2019 and was publicly declared to be fake in June 2019. While it remained active, the profile's manager successfully established contact with 52 people, including representatives of the US Government (a former general and military attaché in Moscow, certain government advisers). The exact purpose of the profile is not clear (spear-phishing tool, gaining access to event agendas to receive credentials or notifications, mapping the network of contacts of the connections, a test, a joke), but what set it apart was the fact that the photo used was synthetically generated, and therefore able to elude the techniques used by social media to verify users' accounts. R. Satter, "Experts: Spy used AI-generated face to connect with targets", AP News, 13/VI/2019.

¹⁹ In military contexts, where additional verification tools are absent, the deception could be perpetrated using a video or audio relaying orders through the chain of command.

of satellite or other images to mislead an observer and influence that person's judgements, decisions, and actions.²⁰

In addition to deception, synthetic media could serve as tools for extorting and coercing individual decision-makers in charge of national security-related functions. In this case, the weaponisation of the media involves the threat of disclosure of the manipulated content, which could result in reputational, image-related, or economic damage for the potential victims. In the attempt to influence the victim, the perpetrator could also leverage the fact that the subsequent reporting of the fake nature of the content disclosed will not protect the victim from short-term damages. Coercion could also take the form of decisions, actions, or interventions to be taken by the victims to respond to the dissemination of altered digital content.

In addition to being weaponised against specific organisations or individuals, synthetic media could also be employed to influence a wider audience. They could be distributed using various on/offline platforms to sway individual cognitive processes and induce a certain behaviour in the target audience. Their use would therefore be intended to mobilise public opinion and thereby exert pressure on political decision-makers, influence their decision-making process, compromise their leadership and, by doing so, obtain political and institutional paralysis, erode citizens' trust in these figures and the institutions they represent, and amplify existing social divisions. The mobilisation of civil society could escalate into major public disorders or, as far as military confrontations are concerned, exacerbate clashes between the parties involved in a conflict (for example, this could involve a deepfake video reporting possible atrocities committed by representatives of disputing factions). In a slightly different context, this was the method used in the attempted coup in Gabon in 2019, which was intended to remove the country's President, Ali Bongo. In that case, the attempted coup was triggered by the broadcasting of a video showing the President in good health, despite various sources of information claiming that he was seriously ill or in fact dead. The video, which has not been proven to be inauthentic, was in any case denounced as a deepfake by opponents of the President and used by them as a pretext for his removal.²¹ It is also possible to imagine cases where deepfake content may be used in operations intended to compromise or undermine the economic and financial stability of a country. This could be a hypothetical scenario in which synthetic media are used to distort the correct operation of financial markets or to damage certain sectors of the economy. A scenario of this kind could occur whereas manipulated content suggesting an imminent crisis in the banking/financial sector would generate a situation of panic and induced individuals to withdraw their money from banks.²²

²⁰ C. Xu and B. Zhao, "Satellite Image Spoofing: Creating Remote Sensing Dataset with Generative Adversarial Networks", 10th International conference on geographic information science (GIScience 2018), pp. 67:1–67:6.

²¹ A. Breland, "The Bizarre and Terrifying Case of the 'Deepfake' Video that Helped Bring an African Nation to the Brink", *MotherJones*, 15/III/2019.

²² A. Katwala, "The Metro Bank hoax shows the immense power of fake news on WhatsApp", *Wired*, May 14, 2019.

Conclusions

The weaponisation of synthetic media that is in principle suitable to raise national security implications would most likely occur within information and influence operations targeting public opinion or predefined social groups.

In the other cases discussed above, it seems reasonable to consider that the potential victims (specific individuals or organisations that have responsibility for national security) are adequately equipped to deal with the threat in question (and therefore have appropriate procedures, technologies and organisational settings to deal with the threat). In other words, the scale of the threat affecting these individuals and organisations will be proportional to the degree of knowledge they have of the phenomenon, the existence of internal systems and processes for verification and validation or –in terms of the possible employment of deepfakes for blackmailing/coercion– the existence of mitigation measures, especially those relying on communication.

It is, rather, the public opinion in general that is the most exposed to, and vulnerable against, sophisticated manipulative interventions. Actually, this is true for many other possible manipulative tactics and not only for those relying on synthetic media. From this point of view, deepfakes represent the ultimate, albeit more sophisticated, tool that can be used to influence the decision-making processes and the perceptions of the members of a community through deception. This means that the threat resulting from the weaponisation of synthetic media should be viewed in a broader context of vulnerabilities and risks, the main elements of which are represented by the very nature of human beings. Absent any intervention, the main vulnerability lies in the susceptibility of individuals to accept certain information passively and without any critical examination, to be inclined to believe things that should be considered not to represent reality, where this would be revealed by a more careful and reasoned examination. Therefore, the fundamental element contributing to determining the scale of the deepfake-associated threat resides not so much in the innovative and sophisticated nature of the technology but, rather, in the way information is accessed, consumed, and shared by individuals.