How to ‘deepfake’ your way military victory

A short overview over ‘deepfakes’ and their implications for modern societies in times of rising international tension

‘Deepfakes’ heavily rely on complex neural networks that need quite a bit of training data in order to create a realistic image, sound or video. This essentially led to a ‘deepfake’ economy in which we are seemingly willingly, but still unknowingly participating. Yet, this ignorance does not mean ‘deepfakes’ cannot be used against us, especially in times of rising international tension.

Jessica Fazekas
Jessica Fazekas has a Bachelor of Education in Psychology & Philosophy as well as History from the University of Vienna and is currently in the Master's teachers program. Additionally, she works part-time for the Austrian Ministry of Education.

‘Deepfakes’ and war

During the last few months, even the most politically adverse people have become familiar with the Ukrainian comedian who portrayed a president only to become one in real life. Most of us have heard many of Volodymyr Zelenskyy’s impassioned speeches as well as smirked when one of his witty remarks struck at the heart of defeatist notions. Yet, this rise to notoriety may put him and his nation in harm's way - here is why.

Russia did not only invade Ukraine in February, but it also simultaneously started a propaganda war. One of their many propagandistic warfare tools are ‘deepfakes’.[1] At this point, you may have a flashback to the first weeks of the war and to the various warning posts on all big social media platforms about possible ‘deepfake’-capitulation-videos of Zelenskyy. Maybe you even recall the rather obvious ‘deepfake’ of him asking his fellow countrymen to cease their defense efforts.[2] Either way, the utilization of this new technology as a weapon is rather alarming.

This is not only due to the fact that ‘deepfakes’ have already been identified as being able to attack the very foundation of democratic societies through their potential for distributing disinformation.[3] It is also due to the heightened stakes of wartime. Disinformation in this context may not only lead to ‘light’ harm to individuals and institutions, but also to the loss of human lives in high numbers. Moreover, this particular kind of disinformation is – unbeknownst to many people - closely connected to individuals’ data traces and behaviors on the web.

‘Deepfakes’ are not that new and like most technologies with questionable usages, they are connected to ‘Reddit’ and porn

‘Deepfakes’ or sometimes ‘cheapfakes’, as the less technology-intensive fakes are being called, are not *that* new.[4] The technological groundwork for ‘deepfakes’ – namely the publication concerning ‘Generative Adversarial Networks’ (GANs) - was laid in 2014,[5] but it first gained the attention of mainstream media in late 2017, when one of the technologies' more questionable applications came to light. As it turned out, it allowed people to project the face of a chosen celebrity onto actors in pornographic videos. A Reddit user with the handle ‘deepfakes’ quickly caught on to this possibility and shared the free software ‘FakeApp’, which allowed other users to create such fake videos. After a media backlash, Reddit decided to ban the ‘deepfake’ community.[6]

Nonetheless, the harm was already done. The handle ‘deepfakes’ became synonymous with the phenomenon and ‘deepfakes’ were here to stay.[7] An ‘ecosystem’ of these videos was already established in 2018 and by 2019 there were approximately 14,678 ‘deepfakes’ in circulation - 13,254 of which were pornographic videos on dedicated ‘deepfake’ porn websites. More troubling, ‘deepfakes’ are not a gender-neutral phenomenon: While non-pornographic ‘deepfakes’ target and harm mostly men, pornographic ‘deepfakes’ exclusively victimize women.[8] This observable pattern of privacy violations and sexism only makes it more troubling that the number of ‘deepfakes’ is – at the current rate – doubling every six months.[9]

Unfortunately, the problematic aspects of this new technology do not end here. ‘Deepfakes’ can be used for the purposeful distribution of disinformation and can undermine the public’s trust in various media platforms as well as democratic institutions.[10] Disinformation – in this context – refers to information that is either wrong, inaccurate or misleading and that has been circulated intentionally by state or non-state actor(s) in order to cause harm or to profit from it.[11]

Although they are connected to ‘fake news’, they are actually – by definition – not the same. That is because ‘fake news’ is not seen as a nuanced enough concept to cover all forms of disinformation and thereby ‘deepfakes’. Additionally, it usually pertains to textual media forms, not audio-visual ones like it is the case with ‘deepfakes’.[12]

These gender and disinformation aspects demonstrate the mutually mediating relation between society and ‘deepfakes’ as well as the interplay of societal and technological values.[13] This makes the current rate of increase of ‘deepfakes’ even more worrisome and raises the question of why the number of ‘deepfakes’ is rising.

It may be simple for users to create ‘deepfakes’, but nothing about the underlying technology can be described as ‘simple’[14]

The exponential rise of ‘deepfakes’ is partly caused by the advance in ‘deepfake’-technology. Nowadays, the technology that is needed to create ‘deepfakes’ is steadily becoming more available and easier to use. Additionally, ‘deepfakes’ are projected to only become more realistic and believable in the coming years.[15]

Although they become easier to generate, the technology behind them is getting more complex. Here is a very basic overview of ‘deepfake’ technology:

‘Deepfakes’ can be seen as synthetic audio-visual media that are made through the alteration or the synthetic creation of pictures, videos and/or audio tracks of faces, bodies and/or voices.[16] Fundamentally, ‘deepfakes’ can be created with the help of machine learning techniques – especially deep learning – or advances in computer graphics, computer vision and image recognition.[17]

To come back to the Reddit user that so kindly provided the name for this new technology: His software used AI-created GANs to project the face of the chosen person onto the protagonist in the pornographic video. GANs do this by combining two different ‘neural networks’.[18]

‘Neural networks’ are based on the functionality of the neural network in the human brain. Like the one in our brain, this ‘mock’ network consists of many neurons (small processing units) that are connected to each other, and each of these connections is weighted. Sensors supply data that serves as input and the interpretation of the aforementioned data is the output. During the training period, these weightings are adjusted as needed. The goal is to match the expected and actual outcomes as closely as possible. Because these networks are multi-layered, they are a form of ‘deep learning’ which in itself is a form of ‘machine learning’.[19]

GANs combine two such networks. One network is the ‘generator’ and the other one is the ‘discriminator’. The sensors of the ‘generator’ are fed data about the chosen person and subsequently generate a new, synthetic data set. In order to determine how authentic the new

data set is, it is assessed by the 'discriminator'. If the 'discriminator' reports the set as fabricated it is sent back to the 'generator'. This back and forth is repeated until the 'discriminator' verifies the data set as genuine.[20] Either way, as soon as the 'deepfake' is no longer detected as being fabricated, it can be published and distributed on various platforms for all to see – that's where the real harm is done.

So how much harm can 'deepfakes' cause? – The answer is: Surprisingly much, in various ways, to many different people and institutions. Ah, and it is getting worse…

Once the 'deepfake' is in circulation there are a number of different categories of harm that can ensure. This article will focus on the harm that would be caused in the aggressed nation. However, most of the following aspects are still troublesome in politically stable and peaceful times.

Generally speaking, this article differentiates between 'harms to individuals' and 'harms to society/institutions'.[21] 'Harms to individuals' will be split into two groups: 'harms to listeners or viewers' of 'deepfakes' and 'harms to the targeted subjects' of the 'deepfakes' in order to get a clearer picture.[22]

Additionally, the categories 'weakened epistemic agency' and 'weakened non-epistemic agency' will be brought up in the discussion. 'Weakened epistemic agency' refers here to people's unawareness of their participation in the 'deepfake' economy. 'Weakened non-epistemic agency' on the other hand is significant in this discussion because 'deepfakes' have the potential to affect people's right to make their own life choices in a free and informed way.[23]

Concerning the aggressed nation and the possible 'harms to listeners or viewers' of 'deepfakes' there are two aspects of note: Deception and intimidation.[24] Deceptions and intimidation, like the usage of 'deepfakes' of the nation's leader surrendering, could influence the citizens to act against their own interests.

In a more nuanced approach, the aggressor could start a targeted deception and intimidation campaign before an open conflict even breaks out. An aggressor could thereby utilize 'deepfakes' and social media bubbles[25] in order to deceive or radicalize parts of the aggressed nation's population into taking their side. This may lead to a change in political leadership that is more favorable for the aggressor, in civil unrest or the foundation of a partisan group(s) that sides with the aggressor in an open conflict.

Such a strategy would also constitute 'harms to the targeted subject' of the 'deepfake' since these kinds of approaches would likely attack the subject's reputation or would misattribute specific notions or statements to the person.[26]

‘Harms to society/institutions’ in such a situation could therefore entail undermining the trust in democratic elections, the manipulation of elections, sabotaging the public discourse of policy issues, lowering trust in political and public institutions, strengthening existing public divides, risking public safety through misleading information, undermining diplomacy (for example by publishing a ‘deepfake’ that shows a diplomate of the aggressed nation badmouth an allied nation), jeopardizing national security (this could be done through ‘deepfaking’ sensitive military information) and disempowering journalism.[27]

‘Deepfakes’ also have the potential to be used to harm the reputation of certain individuals in the aggressor nation like foreign diplomats, members of the opposition parties, human rights activists and even civilians that are openly against the conflict. It is also entirely in the realm of possibility that sufficiently made ‘deepfakes’ could be used to accuse an adverse individual of a crime to ensure the ‘lawful’ elimination of the opposition.

These hypotheticals all exemplify weakened epistemic and non-epistemic agency[28] as well as a debilitation of the epistemic backstop upon which we rely for our testimonial practises.[29] This would strengthen the so-called “liers dividend” – somebody could therefore have been recorded doing something (that the person *actually* did), claim that it is a ‘deepfake’ and be believed.[30]

Whilst all these concerns simultaneously paint a dystopian picture as well as a rather shockingly familiar one,[31] none of these notions refer to the ‘it is getting worse’ part of the headline.

More important than the implications for our testimonial practices are the aforementioned weakened epistemic and non-epistemic agencies, although it might not be obvious at first glance. The reason why the forms of agency are so important in this context is because they constitute the prerequisite and the inevitable outcome of these hypotheticals.

The ‘deepfake’ conundrum: How ignorance can haunt us

To put it more clearly: The most concerning and universal harmful effects of ‘deepfakes’ have nothing to do with elections, privacy issues [32], attacks on our institutions[33] or the epistemic backstop[34], because we know about these issues already and are currently working on them. There are currently a number of governance proposals on national and international levels[35] as well as technical solutions for the detection of ‘deepfakes’[36].

Far more concerning are the behaviors and things that we engage in every day and don’t question at all – the things that are nearest to us and therefore furthest away from our critical minds – that are, in the current absence of solutions – endangering us. Unfortunately, this is especially the case with social media.[37] While some of the consequences of data collection have become known to the general public in the last couple of years, its possibilities for ‘deepfakes’ have not. ‘Deepfakes’ need training data and the more there is, the better. So – mostly unbeknownst to us – social media usage involved a trade: Our data for benefits like

attention. So unbenounced to us, we are actually participating in a ‘deepfake’ economy and it is this unawareness that weakens our epistemic agency.

To come back to the preceding hypotheticals: The fundamental prerequisite of all these cases is the ignorance of the subject of the ‘deepfakes’, because they, at some point in time, willingly, but unwittingly ‘traded’ the training data for these ‘deepfakes’ through social media or other forms of media. This uninformed action can then lead to all kinds of harm, as shown in the aforementioned hypotheticals. Furthermore, it has the potential to turn people into weak non-epistemic agents through the interference with people's right to free and informed life-choices. To give two poignant examples for this:

Firstly, if a ‘deepfake’ video of a politician of an aggressed nation is circulated in which they are endorsing the aggressor, they might be seen as traitors by their fellow citizens as well as by the aggressed nation's allies. This could lead to infringements on their life decisions such as no longer being able to pursue a political career in certain countries, or to safely stay in their own country.

Secondly, if a ‘deepfake’ video of the president of the aggressed nation is shared, in which they are surrendering to the aggressor, the aggressed nations citizens are forced to make a decision regarding their own behavior. Since they are under the impression that their country surrendered to the aggressor, they are neither informed nor free enough to make the decision to surrender or to keep on fighting. They essentially become weakened non-epistemic agents, which can be fatal in times of war.

This brings us back to Zelenskyy: All the attention that politicians get in peace- and in wartime produces training data that can be used to harm them, their citizens, and the democratic institutions they have sworn to serve. At the same time, politicians like Zelenskyy also heavily rely on (social) media attention in order to rally international support for their countries. Similarly, many Ukrainians have taken to social media to inform the international community of their plight.

Therefore, the worst thing about ‘deepfakes’ is that awareness of the trade that people are making would only be a partial solution to the problem, since war-struck countries still depend on (social) media attention. Fortunately, all of us can help with solving this issue by being aware of its very existence and by voting consciously as well as supporting organizations that are dealing with ‘deepfakes’.

References

[1] Milmo, D., & Sauer, P. (2022, March 19). Deepfakes v pre-bunking: Is Russia losing the infowar? *The Guardian*. https://www.theguardian.com/world/2022/mar/19/russia-ukraine-infowar-deepfakes
[2] Metz, R. (2022, March 25). Deepfakes are now trying to change the course of war. *CNN*. https://edition.cnn.com/2022/03/25/tech/deepfakes-disinformation-war/index.html
[3] Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*, 1753–1820.

[4] Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, *23*, 2072–2098. https://doi.org/10.1177/1461444820925811
[5] Pawelec, M., & Bieß, C. (2021). *Deepfakes. Technikfolgen und Regulierungsfragen aus ethischer und sozialwissenschaftlicher Perspektive*. Nomos Verlagsgesellschaft, 24-37.
[6] Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers' Imprint*, *20*, 1–16.
[7] (Rini, 2020)
[8] Ajder, H., Patrini, G., Cavalli, F. & Cullen, L. (2017, September). Deepstate. The state of deepfakes. Landscape, threats, and impact. *Sensity*. https://sensity.ai/reports/
[9] Sensity Team. (2021, December 21). How to detect a deepfake online: Image forensics and analysis of deepfake videos. *Sensity*. https://sensity.ai/blog/deepfake-detection/how-to-detect-a-deepfake/
[10] (Diakopoulos & Johnson, 2021)
[11] HLEG. (2018). *A multi-dimensional approach to disinformation: Report of the independent High Level Group on fake news and online disinformation.* Publications Office of the European Union. https://hdl.handle.net/1814/70297
[12] (Pawelec & Bieß, 2021, 22-23)
[13] Reijers, W. (2022). Ethics of emerging technologies – Handout 4.
[14] For a more detailed explanation please see also: Mirsky, Y., & Lee, W. (2022). The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, *54*, 1–41. https://doi.org/10.1145/3425780
[15] (Diakopoulos & Johnson, 2021)
[16] (Pawelec & Bieß, 2021, 23)
[17] (Pawelec & Bieß, 2021, 37)
[18] (Pawelec & Bieß, 2021, 37)
[19] (Pawelec & Bieß, 2021, 21-37)
[20] (Pawelec & Bieß, 2021, 37)
[21] (Chesney & Citron, 2019)
[22] (Diakopoulos & Johnson, 2021)
[23] Pham, A., & Castro, C. (2019). The moral limits of the market. The case of consumer scoring data. *Ethics and Information Technology*, *21*, 117–126. https://doi.org/10.1007/s10676-019-09500-7
[24] (Diakopoulos & Johnson, 2021)
[25] Social media bubbles or filter bubbles are created through algorithms and create an online environment in which a user is only confronted by one-sided – often times rather radical – information.
[26] (Diakopoulos & Johnson, 2021)
[27] (Chesney & Citron, 2019)
[28] (Pham & Castro, 2019)
[29] (Rini, 2020)
[30] (Chesney & Citron, 2019)
[31] AN: Which was actually not intended by the author. The above-mentioned scenarios were intended to be hypotheticals.
[32] (Diakopoulos & Johnson, 2021)
[33] (Chesney & Citron, 2019)
[34] (Rini, 2020)
[35] (Pawelec & Bieß, 2021, 174-185)
[36] (Mirsky & Lee, 2022)
[37] (Reijers, 2022)