# Unveiling Deepfakes: Individuals' Affect, Societal Impact and Ethics

**Barto Radman**
Department of Cognitive Science and Artificial Intelligence
Tilburg University
b.radman@tilburguniversity.edu

## Abstract

Deepfakes are a generative AI tool used to create or alter audiovisual content. In a society widely dependent on information technology, such technology has enormous potential for misuse. This discussion paper focuses on affective computing, a field of AI examining human emotions, and generative AI nature of deepfakes. Misinformation is first outlined, leading to the ways that deepfakes alter its already complicated nature. Current developments in deepfakes, along with the process of their creation are then discussed. Impact on individuals, the society and ethics is considered in congruence as essential discussion points. Lastly current methods in detection of deepfake content is shortly discussed. It is expected that initiating a conversation along with gaining an understanding regarding deepfakes will provide the best means to mitigate the negative consequences.

## 1 Introduction

Developments in technology, recently generative Artificial Intelligence (AI), have raised nuanced issues as the public becomes more dependent on their use. AI systems are being developed and deployed on a wide range of tasks by various agents. From individuals, large organizations, and governmental agencies - these systems are a critical part of the information infrastructure (de Ruiter, 2021). It is therefore paramount to address some of the prominent issues within the field, as raising questions and promoting discussion will better ensure safe and beneficial use of the technology. The present paper therefore aims to discuss deepfakes, a generative AI technology used to manipulate audiovisual content (Albahar et al., 2019).

### 1.1 The Nature of Misinformation

Understanding the different facets of misinformation is a crucial first step in evaluating potential implications of deepfake technology. Misinformation is broadly defined as ”false knowledge shared by someone who believes it to be true” (Nour et al., 2022). Distinguishing between facts and false beliefs is a widespread challenge in a society overflowing with information. Technological advances further complicate the nature of misinformation. As a result, misinformation adopts a new feature in which false information is shared with full knowledge of its falsity, deliberately intending to deceive or harm (Wardle & Derakhshan., 2017; Nour et al., 2022). Evidently, it becomes important to understand whether the information is created with particular (potentially malicious) intent or not. Aïmeur et al. (2023) provided a complete review of different aspects of misinformation. Generative AI technology, used purposefully to create misinformation, will pose unseen threats to what constitutes a true fact. Currently, the recognition of such content through AI algorithms and human detection is possible through technical and theoretical understanding (Nour et al., 2022). Nevertheless, such misinformation might eventually become indistinguishable from reality. And when something indistinguishable from reality can influence societal structure, it becomes clear why a grasp of its implications will be the key to fostering benefit and mitigating harm.

### 1.2 Deepfakes

The term deepfakes specifically refers to generative machine learning algorithms which are capable of artificially generating and manipulating audiovisual media (Hosler et al., 2021). Similar technology was initially used to control and mimic facial features of a target person by inserting facial expression of a different person (de Ruiter, 2021). Early algorithms were primitive in the way that it was clearly discernible that the content has been manipulated in some way. Recently, with the rise of deep neural networks, the creation of less discernible content has been made possible. Neural network architectures, such as Generative Adversarial Networks (GANs), can be trained with a large amount of data to learn human facial features. One network is then tasked to generate new content from the learned examples, while a different network attempts to discern whether the new generated content is convincing or not (de Ruiter, 2021). The use of GANs allows the deepfake created content to be continuously refined. If you consider the novel use of GANs, and attach the societal context of information distribution predominantly through platforms consisting of user generated content, the danger (and potential) of deepfakes becomes clear (Veerasamy & Pieterse, 2022). What follows are some prominent points of concerns of deepfake technology, starting with implications for individuals, moving to broader societal contexts and ultimately ethics. The last section will focus on some of the technical issues regarding detection of deepfake content.

## 2 Key issues with deepfakes

Deepfakes already have an immense potential to influence the data driven world that we find ourselves in (Albahar et al., 2019). This section will first discuss potential key issues for the individual experiencing these rapid improvements in generative AI. (i) It will explore potential impacts of agents becoming evermore capable of affective computing. Artificial agents, created through the use of deepfake technology, will be able to effectively display emotions, artificial empathy and artificial moral sensitivity. The second section will deal with broader societal impacts of deepfakes. (ii) Specifically, the intricate issue of creation and distribution of fake news, along with detection and fact checking thereof. Lastly, the impact of generative AI on the sphere of ethics will be considered. (iii) The distinct ethical challenges, potential shift in what constitutes consent and responsibility for the harm created through the use of deepfakes.(iv) Lastly, some of the issues and potential in successfully detecting deepfakes will be elaborated. The motivation and hope of this discussion paper is that by leading a fruitful discussion and promoting understanding, the misuse of the technology can be mitigated. The key to safe use of generative AI technology, in any form, will ultimately be a widespread awareness and interest in these topics.

### 2.1 Individuals, affective computing and emotional manipulation

Affective computing is an active area of research in artificial intelligence. The research mainly focuses on artificial agents specifically trained to express particular cognitive functions found in humans (Picard, 2000). It covers various aspects of affect, such as emotion expression, communication, and recognition, along with physically embodied agents. Picard (2000) outlines prominent challenges in affective computing, namely the fact that progress is limited due to the nature of emotions and potential for emotional manipulation. People themselves have difficulty in attributing emotional states to themselves or others, while emotions themselves are often thought to be private in nature (Picard, 2000). Progress in the field of affective computing has been achieved through artificial agents learning from real human data. In similar fashion, deepfake models learn by learning particular features of a person in order to be able to create or alter content. Research by López-Gil et al. (2022) reports that deepfake algorithms are still not capable of generation and expression of novel (more complex) emotions, likely due to lack of input data which convey emotions clearly. The emotion generation is unnaturally displayed, thus failing to appeal to the semantic features displayed by humans (Holser, 2021). In a sense, the current models learn affect too rigorously, not accounting for irregularities, mismatch and the dynamic aspects of emotion often present in humans. People are innately well trained to distinguish whether emotions are genuine or purely imitated. Despite the limitations on high level semantic features of emotion, it is critical to recognize that the technology is still in its infamy and research in deepfake emotion recognition is still scarce. An individual interacting with a model trained purposefully to imitate human affect in order to achieve an aim presents a challenge. For example, Yang et al. (2022) reported successful use of deepfakes for development of artificial empathy in clinical settings through interactions between patients. It remains to be seen whether the impact of deepfakes on individuals moral sensitivity, attention, motivation and emotion manipulation is significant (Ramachandran, 2023). Long term exposure to more realistic deepfake content is not yet known. In particular, emotional manipulation could be one of the first areas of interest when using deepfakes for malicious purposes. This potential, along with improvements of the technology, motivates more research on the topic while the models are still rather primitive. Holser (2021) provided a functioning framework for evaluation of deepfake content focusing on ”non natural and inconsistent emotions”, while other research proposed the possibility of models assisted by human evaluators to distinguish artificial content (Masood et al., 2022). It is evident that the technology is rapidly evolving to overcome some of the mentioned limitations (Masood et al., 2022). Thus, the discussion about potential danger to individuals should be made explicit, along with of the risks associated with refinement of deepfake technology. A more prominent and already successful misuse of deepfakes can be found in affecting the society through fake news.

### 2.2 Societal implications of fake news and the detection thereof

Currently most dangerous misuse of deepfake technology can be found in the media industry. Media has long held the responsibility for communicating important, truthful and objective information (Gálik, 2019). While possessing own inherent biases, individuals generally posses enough understanding to discern media outlets which are spreading misinformation. Today, both truthful and malicious information is more easily spread than ever due to the increasing societal dependency on digital information systems. An example of such malicious information is fake news. Fake news encompasses a wide range of disinformation, usually created to resemble credible news while intentionally spreading misleading information (Albahar et al., 2019). A survey by Botha et al., (2020) has reported that fake news are already a significant problem globally. Influencing public opinion and political agendas (Vizoso et al., 2021), inspiring riots and spreading information promoting hate or violence (Botha et al., 2020) - are just some of the possible consequences of fake news. It is therefore no wonder that the governments, along with large tech corporations are setting up projects to deal with misinformation spreading over the immensely connected society (Vizoso et al., 2021). Deepfakes again pose a novel feature to an existing issue. Fake news are now not required to be created manually, but through the use of deepfake technology this can be achieved automatically and at a much faster rate. Being able to alter or create new content easily with any intent is an arduous issue to deal with. Examples of deepfake disinformation being spread can be found during the 2016 US presidential election (Botha et al., 2020), various propaganda and spreading of fake scientific information regarding climate change or recycling. Such falsities are not just problematic in swaying the public opinion through emotionally manipulation, but also poses a threat to the credibility of journalism (Vizoso et al., 2021). The difficulties with which false information is counteracted today also harms the reputation of truthful and objective information news sharing platforms. Scientific publications, often coupled with image evidence, are also at risk of researchers using deepfakes to fabricate their evidence. Evidence fabricated by deepfakes could result in false conclusions, practices and harm to the credibility of science. With the future depending on factual information being provided by journalists and the scientific community, deepfakes will play a pivotal role. It is becoming increasingly more difficult for the society to gather truthful information due to most of the news being acquired through social media, a digital playground for spreading deepfake disinformation. Quality of fakes constantly increases, seemingly causing a tug-o-war between production and detection.

## 2.3 Distinct wrong of deepfakes and responsibility

The issue of the wider societal impact of deepfakes, especially on fake news and the credibility of journalism, raises the discussion to an even wider front. The sphere of possibilities for misrepresentation, morality, distinct wrongdoing and responsibility rapidly increases with the expected advancements of deepfake technology. The crux of the ethical consideration, and the evaluation thereof, lies in the vast amount of different reasons for which deepfake technology can be used (Albahar et al., 2019). As discussed in the previous sections, deepfakes can be used to individually manipulate through creation of realistic affective features for which humans are innately programmed to respond in a way which can be used to emotionally manipulate. In a similar way, the societal impact of realistic and automatically generated fake news, disinformation and loss of journalistic credibility will have a considerable impact on the world. Novel technologies often contain benign, beneficial and negative consequences depending on their use. What logically follows is a necessity for discussion on whether such technology is distinctly wrong, and if so, under which outlined conditions (de Ruiter, 2021). It therefore does not come as a surprise that deepfakes, which will be able to (or already can) effectively portray people, will evoke a sense of moral wrong (de Ruiter, 2021). First, deepfakes are developed and deployed for a variety of beneficial uses such as regeneration of voices, reducing costs of film production and empathy in clinical settings. A common feature among these uses is that the people portrayed, along with the manipulation of the content is done with consent of the targets. They represent individuals or collectives in a way that is not malicious. Viewers are not deceived by this kind of content, but the grounded perception is simply altered or recreated truthfully (de Ruiter, 2021). While it can be argued that any kind of artificial manipulation of reality is ethically wrong, the intent upon which the content should be taken into account. On the other hand there is undoubtedly even more space for malicious compared to beneficial use. Politics, false representations, non-consensual content, and cyber terrorism are just a few examples which are not often discussed. Deepfakes could even be used to provide false evidence in court, inspire havoc in societal and domestic relations, and invoke violence. In such ”negative” uses of the technology, the individuals or the collective doesn’t have the necessary agency to decide whether the representation is consensual or damaging, but the agents using the technology posses all of the agency (Ramachandran et al., 2023). The scope of possible deontological ethical considerations of deepfakes is a complex issue, however it is evident that the possibility of wrongdoing might outweigh the beneficial uses. In such cases who is responsible for the damage? It could be the person creating the algorithm for the manipulation of the specific content, a person utilizing the algorithm, the media outlets promoting the content or the targeted people in which potentially negative reactions to the fake content are evoked. These answers to such questions might not be clear and could not have evident answers in current law systems. What is more clear is that deepfakes do not posses a clear distinct moral wrong, but should be held under suspicion depending on their use (de Ruiter, 2021). With the technology presumably still in its infamy, like with many other generative AI technologies, the questions of ethics along with other discussed implications should be addressed prior to the wide deployment. The last section will delve into some of the detection techniques which could be used to make answers to the key questions more manageable.

## 2.4 Detection of deepfakes

Detection of deepfakes is already a prominent area of research. Systematic review conducted by Vizoso et al. (2021) reported that media sources such as *The wall Street Journal*, *The Washington Post* and *Reuters* have divisions of non connected journalists with an objective of detecting deepfake information. Professional journalists with experience in the ways that the media is usually created poses a unique ability to discern what is clearly fake. With deepfakes still struggling with particular high level semantic facial features (Holser et al., 2021), it is also possible for individuals interacting with the content to clearly discern, or simply know that something might be fake. But what happens when the technology is able to advance the realism of the content to the point that the humans can not discern it themselves. Tech corporations such as Google and X (formerly Twitter), collaborate with defense agencies in order to develop AI tools for algorithmically detecting deepfakes through features which might not be salient to the human eye. There are initiatives to provide more widespread ”noisy” data in order to fight against deepfakes, such as in the case of artists corrupting AI generated art. This noisy data would be available in datasets which deepfake algorithms use for training. When training on such data, the deepfakes would incorporate the noise, thus making them less realistic and easily discernible. Apart from human and AI detection, the created misinformation is identified and propertly labeled by large social media corporations (Vizoso et al., 2021). Investing in and continually keeping up with the capabilities of generative AI technologies will undoubtedly be a challenge. A more detailed overview of current developments in specific detection methods of deepfakes can be found in the review article by Masood et al. (2022).

## 3 Conclusion

Deepfakes will pose unseen challenges for individuals, society and the nature of ethics. A creation, or realistic manipulation, of content for a wide variety of purposes requires awareness and understanding of the topic. People could be individually emotionally targeted, fake news could be wide spread and the ethical considerations along with law regulations will have to keep up with the rapid evolution of the technology. This discussion paper has outlined notable issues on all of these levels. Being able to detect deepfakes through the use of human evaluators, along with AI algorithms provides a means to combat some of these issues. The proponents of the technology advocate for the vast beneficial uses, while skeptics believe that the misuse far out weights the benefits. It is therefore left to be seen in which way the technology could further develop, and it is up to governments, large corporations and individuals to remain skeptical and aware of the possible influences. It is important to advocate and promote safe use while the technology remains in its infamy. While this paper has outlined some of the prominent issues, novel issues rapidly arise, along with novel solutions. Investments made in research on the current deepfake issues along with possible future ones will be the key to harnessing the greatest benefit that the technology has to provide.

## References

[Albahar and Almalki, 2019] Marwan Albahar and Jameel Almalki. Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, 97:22, 2019.

[Aïmeur *et al.*, 2023] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13, 12 2023.

[Botha and Pieterse, 2020] J Botha and H Pieterse. Fake news and deepfakes: A dangerous threat for 21 st century information security, 2020.

[de Ruiter, 2021] Adrienne de Ruiter. The distinct wrong of deepfakes. *Philosophy and Technology*, 34:1311–1332, 12 2021.

[Gálik, 2019] Slavomír Gálik. On ontological definition of media truth and the role of media, 2019.

[Hosler *et al.*, 2021] Brian Hosler, Davide Salvi, Anthony Murray, Fabio Antonacci, Paolo Bestagini, Stefano Tubaro, and Matthew C. Stamm. Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1013–1022, 2021.

[López-Gil *et al.*, 2022] Juan Miguel López-Gil, Rosa Gil, and Roberto García. Do deepfakes adequately display emotions? a study on deepfake facial emotion expression. *Computational Intelligence and Neuroscience*, 2022, 2022.

[Masood *et al.*, 2023] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, 53:3974–4026, 2023.

[Nour and Gelfand, 2022] Nika Nour and Julia Gelfand. Insights and issues that challenge and demonstrate the role of gl deepfakes: A digital transformation leads to misinformation, 2022.

[Ramachandran *et al.*, 2023] V Ramachandran, C Hardebolle, N Kotluk, T Ebrahimi, R Riedl, P Jermann, and R Tormey. A multimodal measurement of the impact of deepfakes on the a multimodal measurement of the impact of deepfakes on the ethical reasoning and affective reactions of students. 2023.

[Veerasamy and Pieterse, 2022] Namosha Veerasamy and Heloise Pieterse. Rising above misinformation and deepfakes, 2022.

[Yang *et al.*, 2022] Hsuan Chia Yang, Annisa Ristya Rahmanti, Chih Wei Huang, and Yu Chuan Jack Li. How can research on artificial empathy be enhanced by applying deepfakes? *Journal of Medical Internet Research*, 24, 3 2022.

[Ángel Vizoso *et al.*, 2021] Ángel Vizoso, Martín Vazálvarez, and Xosé López-García. Fighting deepfakes: Media and internet giants' converging and diverging strategies against hi-tech misinformation. *Media and Communication*, 9:291–300, 2021.