The Blurred Lines Between Reality & Fiction

By: Daphne Tang

EMPLID: 24097532

Email: dtang000@citymail.cuny.edu

In my research paper on AI deepfake technology, I briefly introduce the term “deepfake,” its definition, and its origin. I will explain the significance of this study as it raises ethical concerns and potential risks in cybersecurity. I will give an overview of the historical development, from the early methods of deepfakes to recent advancements. I provide various examples of real-world applications of AI deepfakes. I will explain the common methods used in its creation, including generative adversarial networks, autoencoders, and face-swapping technologies and how they work. I identify the ethical concerns and limitations, as well as showcase several case studies to highlight its impact on our society. I explore possible future trends and implications of societal and cultural shifts due to deepfake technology.

Deepfakes are defined as “an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said,” [3] according to the Merriam-Webster dictionary. The term derives from "deep learning" and "fake." Deep learning refers to machine learning using artificial neural networks with multiple layers of algorithms. “Fake” refers to the unreal media where a person is realistically manipulated or replaced using artificial intelligence (AI).[3] With remarkable advancement in this transformative technology, it is now possible to create hyper-realistic videos and images in which it is difficult to distinguish from real content. This technology carries substantial societal, political, and ethical implications. The ability to forge such convincing visual and auditory content has sparked concerns about misinformation, erosion of trust, and the potential misuse of manipulated media. Additionally, this affects cybersecurity, as well as authentication and verification processes.

The term “deepfake” was originally used by a Reddit user in late 2017, who created an online news and aggregation site to share pornographic videos, utilizing open-source face-swapping technology.[7] Historically, people attempted to manipulate media through rudimentary techniques, such as basic image and video manipulation techniques. However, in recent years, there have been significant developments due to the use of deep learning. The development of Generative Adversarial Networks (GANs) and advanced face-swapping algorithms are especially noteworthy as they contribute to the unexpected realism achievable in deepfakes. These advancements can be used in various ways including but not limited to the entertainment industry, enabling seamless visual effects in movies and television. However, it also introduced serious concerns in political manipulation, as they can fabricate speeches and manipulate political discourse. [6] Deepfakes can also be used to commit malicious acts like identity theft, misinformation, and cyber-espionage. With the rise of deepfake technology, we must consider countermeasures against such misusage. We must understand the usage and application of AI deepfakes, with an exploration of their methods, challenges, and societal implications.

The evolution of AI deepfakes thrives with sophisticated techniques, similarly to Generative Adversarial Networks (GANs). GANs use deep learning methods, such as classes of neural networks, for generative modeling. Generative modeling is an unsupervised machine learning method that automatically discovers and learns the patterns of input data to generate and output new examples, drawing from the original dataset. [1] There are two interconnected network for the two sub-models: a generator and a discriminator. The generator generates new data or examples in the form of images or videos, while the discriminator learns to differentiate between real (from the domain) and fake (generated) content. [1] The models train together until the discriminator model is fooled around half the time, which indicates the generator model has fabricated some convincing examples. [1] This helps GANs create surprisingly authentic deepfakes by repeatedly refining the generated output.

Autoencoders are another powerful method in producing deepfakes. Autoencoders utilize the encoding-decoding principle by compressing input data into a compact resizing and reconstructing it. [5] For creating deepfakes, autoencoders learn latent facial features during training. During generation, the decoders are swapped, for example, original face A is transformed by decoder of face B to generate face A with similar facial features of face B. This allows for realistic facial manipulations. [5] Another technique is the face-swapping technology, which combines both traditional and deep learning-based approaches. Traditional methods depend on the manual mapping of facial features, while deep learning-based techniques utilize neural networks to automatically assimilate and transform facial expressions. [8] The combination of these methods clearly shows the multifaceted landscape of AI deepfake creation, in which neural networks and advanced algorithms play a significant role in.

As the evolution of AI deepfakes continues, there is a rising concern that humans may no longer be able to distinguish between real and fake content. Typically, one can differentiate between genuine content and deepfakes by paying deep attention to one's face. For example, asking yourself "Is the person blinking too much or too little? Do their eyebrows fit their face? Is the person's hair in the wrong spot? Does their skin look airbrushed, or are there too many wrinkles?" [7] One can also determine such by examining if the audio matches the appearance. Another method is examining the lighting for some form of reflection shown from a person's glasses under the light, as deepfakes usually fail to perfectly represent the natural physics of lighting. [7] Advancements in machine learning algorithms have aided the development of detection tools. However, deepfake technology is often more advanced than these efforts. There are many challenges in developing effective detection techniques as it must account for the fast evolution of deepfake methods, the massive amounts of digital content generated daily, and the diverse platforms hosting all the content. There are also many ethical concerns of using AI deepfakes, as the potential misuse of this technology range from political manipulation to identity theft. As deepfakes blur the lines between reality and fiction, there is a spark of distrust in media, institutions, and interpersonal relationships. This emphasizes the need to address both technical and ethical concerns.

There are many instances in which AI deepfakes caused real-world impact that emphasize how urgent it is to understand and address this technology. One instance is the manipulation of political figures through deepfakes. These deepfakes of political figures are orchestrated to distort public perception of the figures and sway opinions. There are videos of political figures making fabricated statements and in fictional scenarios. [6] This highlights the potential for the misuse of this technology to disrupt democratic processes and encourage public discourse. Moreover, there is also a surge of deepfake incidents in the entertainment industry involving celebrities. MrBeast, one of the world's biggest YouTuber, is a content creator known to give away large amounts of his money to others. He was recently used in a deepfake video intended to scam innocent viewers online, in which "MrBeast" offers viewers brand new iPhones for only $2.[4] A Meta spokesperson stated that, "We don't allow this kind of content on our platforms and have removed it. We're constantly working to improve our systems and encourage anyone who sees content they believe breaks our rules to report it using our in-app tools so we can investigate and take action." [4] It is reassuring to know that those who host these platforms are wary of the rising concerns of AI deepfakes. However, this means the public must stay vigilant and cannot trust all the media released on these platforms as well. There are also fabricated interviews to manipulate movie scenes, raising questions about the validity of digital content and the potential exploitation of public figures.[2] These case studies not only showcase the versatility of deepfake applications but also emphasize the need for improvements in safeguards to mitigate manipulated media on political landscapes and cultural perceptions.

AI deepfake technology's future trends indicate more sophisticated technologies in deepfake development. As AI continues to advance, generative models will be refined, the data sources will be integrated, and audio-visual manipulation technologies will enhance the realism and potential applications of deepfakes. The consequences of this technological evolution will affect societal and cultural aspects of our lives, as there is much potential for deepfakes to influence public opinion and reshape cultural narratives. It may even impact interpersonal trust, raising questions about the long-term implications on society. Cybersecurity remains as a critical problem if deepfakes can potentially be used to against authentication systems, manipulate financial transactions, or conduct cyber-espionage. This ongoing evolution of this technology calls for increased vigilance in cybersecurity methods. We must be wary of these future trends and understand how it may impact our society and how cybersecurity must strengthen to not be deceived by AI deepfake technology.

In conclusion, this research paper has explored the topic of AI deepfakes, exploring their historical development, real-world applications, methods of creation, challenges and concerns it poses, case studies, and future trends. The rise of Generative Adversarial Networks (GANs), autoencoders, and face-swapping techniques is shown to be evolutionary in how they are used effectively in deepfake creation. Challenges and ethical concerns continue as we lack accurate detection methods, encouraging the development of advanced detection technologies. Case studies involving political figures and celebrities highlight the impact of deepfakes on public perception and societal trust. We must not undermine the importance future trends, potential societal shifts, and cybersecurity challenges posed by the evolution of deepfake technology. To face these challenges, we must navigate the complex implications and ethical concerns surrounding the ever-evolving topic of AI deepfakes.

Works Cited

1. V Brownlee, Jason. "A Gentle Introduction to Generative Adversarial Networks (GANs)" https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/
2. "Can Congress Stop Celebrity Deepfakes on Social Media?" Wall Street Journal. https://www.wsj.com/podcasts/google-news-update/can-congress-stop-celebrity-deepfakes-on-social-media/dbd3dfe8-6848-427d-a766-290d3d40c162
3. V "Deepfake." Merriam-Webster.com Dictionary, Merriam-Webster, https://www.merriam-webster.com/dictionary/deepfake. Accessed 14 Nov. 2023.
4. V Gerken, Tom. "MrBeast and BBC stars used in deepfake scam videos" https://www.bbc.com/news/technology-66993651
5. V Masood, Momina & Nawaz, Marriam & Malik, Khalid & Javed, Ali & Irtaza, Aun. (2021). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward.
6. V McKenzie, Bryan, "Is That Real? Deepfakes Could Pose Danger to Free Elections" https://news.virginia.edu/content/real-deepfakes-could-pose-danger-free-elections
7. V Somers, Meredith. "Deepfakes, explained." https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained
8. V Walczyna, Tomasz. "Quick Overview of Face Swap Deep Fakes" https://www.mdpi.com/2076-3417/13/11/6711