Afoma Muojekwu

Professor Xiaojing Liao

INFO-I433

05 May 2025

Project Report

Media is one of the biggest sources of connection, news, social networking, communication, entertainment, and expansion. The content displayed on platforms like TikTok, Instagram, X, Facebook, etc., gain a lot of traction and therefore make up the majority of the content which people are taking in. With impact comes responsibility. Some content shown to the masses is honest and well-intentioned, while others are under the guise of being authentic when they are not. This is the worrisome part about AI. It has great capability to do great things and take the world in a forward direction like never before with its innovation, but it has the power to do the exact opposite as well. There are many ways for written text, images, audios, and videos to be inauthentic, incorrect, or harmful to others.

As a start, I believe it is integral that some key terms are defined to serve as a precursor for our understanding. The first term that must be defined is real. For something to be real, it would need to be actually existing as a thing or occurring in fact; not imagined or supposed. The next term is fake. For something to be fake, it needs to be a thing that is not genuine or in other words, a forgery or a sham. Another term to define would be deep. In order for something to be deep, it needs to be very intense or extreme. All this leads us into our final and most important term; deepfake. A deepfake is officially defined as “a video or image of a person in which their face or body has been digitally altered so that they appear to be someone else, typically used maliciously or to spread false information” (University of Virginia 2023). This term is the catalyst of everything presented past this point. For this project, I would like to focus on the misuse of generative AI through deepfakes.

To begin, I plan to discuss my research on the misuse of AI deepfakes by finding some real-life examples ranging from altered images to videos of political and celebrity figures saying and doing things that they never truly did and the significance of this in section one. From here, I will explore the spread of misinformation in important discussions using instances that have occurred already to show how deepfakes take it all a step further in section two. I will survey the different perspectives on deepfakes, both negative and positive, and the reasoning behind them in section three before moving into section four to discuss the tools and resources used to combat inauthentic or harmful AI from a broad perspective. In the final section, section five, I will go into detail about what the existing tools could contribute to the ultimate solution, as well as go over some suggestions of my own after having researched the topic.

As an example of a well-intentioned deepfake, there exists a video of Lebron James teaching people how to do math. Specifically, how to understand limits and the purpose that understanding these limits serve. This could be beneficial because it allows a topic that may be seen as difficult to understand for some people or difficult to teach by some educators to be taught in a fun way. By using a celebrity that is known for being humorous and talented, people are more inclined to hang on to every word being said. So if Lebron James is speaking about limits, it may appeal to the youth more and the concept of limits may be communicated in a more effective way. While the video holds no bad intentions to it, Lebron James never consented to his likeness, voice and image being used for media content in which he is saying things that never came out of his mouth. So although this video is a positive deepfake, it is still not in alignment with the Deepfakes Accountability Act or the ELVIS Act, therefore it could qualify as malicious deepfake content.

Another example, this time of an ill-intentioned deepfake, would be a video of President Zelensky of Ukraine ordering his soldiers to surrender and put down their arms. This is very ill intentioned and malicious because it is giving war orders to those fighting the war from an important political figure, and if believed, could cause a blip in time where adversaries could take advantage of that vulnerability before being recognized. This type of deepfake could contribute to violence, loss of life, and could even be a catalyst for an entire country being taken over. Luckily, the deepfake was so odd that not many people took it as truth, and it did not end up having these dire consequences. Either way, that specific deepfake is surely not in alignment with the Deepfakes Accountability Act or the ELVIS Act and does qualify as malicious deepfake content.

A couple of resources that have contributed to the mitigate the proliferation of harmful AI generated deepfakes include the “Deepware Scanner, DuckDuckGoose, and Google SynthID” (Harris 2024). If you are a frequent user of YouTube, like many people, it is likely that you may have come across the first platform; Deepware Scanner. Deepware Scanner helps mitigate the spread of generative artificial intelligence “content that includes human faces” (Owda 2024). All you need is the URL for the video you’re wanting to test the validity of. Afterwards, you run the URL through the scanner and use the report ID to get the results. The second tool, DuckDuckGoose, takes it a step further, serving as a hub for six different deepfake detection tools. This solution detects image, video, and audio manipulation in real-time. Not only does it detect, but it also offers a breakdown of which elements it used to determine the ingenuity of the content. The final tool, Google Synth ID, is a watermarking tool that allows AI content creators to add digital watermarks to their work. This signals to others that AI in that specific instance is to be trusted and does not participate in the wave of negative intentions we often see in modern AI today.

The next facet of the production of existing solutions that have already come to light regarding deepfakes would be law. Through the law, there has been steps to reduce the use of people’s content without their consent through the acts being proposed. A couple of these acts include the Deepfakes Accountability Act and the ELVIS Act. The Deepfakes Accountability Act was “sponsored by Rep. Clarke, Yvette D. of New York and introduced on September 20th of 2023 through the subcommittee on Emergency Management and Technology” (Library of Congress 2023). It is cited as the defending of “each and every person from false appearances by keeping exploitation subject to accountability” (Library of Congress 2023). The ELVIS Act made in Tennessee and signed by Governor Bill Lee in 2024 is an act that protects from the “non-consensual use of cloning likeness and voice” (Library of Congress 2024). These laws can help demotivate the generation of deepfakes due to the legal trouble that could follow after it. There has also been major discussion of limiting the use of deepfakes in political campaigns due to the risk of election interference and political persuasion. A couple of states that have shown to be proactive in tackling this issue would be California, Texas, and Tennessee.

Some suggestions that I would propose on this issue from a user standpoint would be to exercise caution through being attentive to privacy. It is important that we redirect our focus to privacy and ensure that we “limit other users’ access to our sensitive content” (Forbes 2024). The less access, the less chance of someone with bad intentions causing harm. The next suggestion would be utilizing the technique of digital watermarking. If your content has an official stamp on it, you can take legal action to further protect yourself if that legality is breached. Without this layer of protection over your content, you could report malicious behavior as it is, but this watermarking makes it all the more effective.

To go along with this, I would recommend utilizing the existing tools as well. Google SynthID, DuckDuckGoose, and Deepware Scanner serves as an amazing defense mechanism against malicious deepfakes and serves as a great teacher to “help users spot what makes it a deepfake” to better protect themselves in the long run (AI for Education 2025). It is important on a company level to incorporate a good amount of education on the topic of deepfakes to its team members for this same purpose. This becomes even more integral as we move from individual to company, because a company contains sensitive information on each individual in the company as well as all of its clients. It is a company’s moral responsibility to keep their client and staff information confidential, and when a security breach occurs on this level, it could cause a major loss of trust between company and client, as well as employer and employee.

Finally, when malicious behavior is detected, it is best to report it right away. To limit your online footprint is helpful, but it is not an absolute failsafe. Reporting in conjunction with limiting your online footprint as a joint effort makes it so that the sooner and more frequently something malicious gets reported, the more subdued the aftermath of its harm can end up being. The less people who have to see the content, the less proliferation of that content. While keeping these suggestions in mind, it is important to remember the three tiers of the CIA Triad: confidentiality, integrity, and accessibility. It is hard to have all three for a user while trying to mitigate deepfakes, but if there is a way to execute these suggestions while still striking a balance between all three, it would be the perfect add on to the already existing solutions.

As aforementioned, we know that Media is one of the biggest sources of connection, news, social networking, communication, entertainment, and expansion. Those who take advantage of this should be aware of this reality and strive to use the internet honestly. Unfortunately this is not always the case. Through this report, we have discovered some existing tools like DuckDuckGoose, Deepware Scanner, and Google SynthID as well as some laws that have already been proposed to serve as a solution to the proliferation of deepfakes. We went over a couple of suggestions such as improving our privacy and digital footprint online, reporting malicious or suspicious behavior, digital marketing, increasing education for awareness, and making use of the existing Deepware detection apps.

On a philosophical level, I believe it is sad to see something that could be so powerful and effective in bringing the human race better be used for evil. Human beings will make whatever they can out of things, no matter where we inhabit. We inhabit the earth, and some human beings do litter, and other human beings water plants. If we inhabit mars, some human beings will litter on mars, and others will tend to the planet in a way that is beneficial. Either way, it is not the earth’s fault or mars’ fault if things go awry. Any place we inhabit is a reflection of us. The cybersphere is no different. Cybercrime on the internet reflects crime in the real world. Overall, we want to keep the internet accessible to citizens on the cybersphere while also keeping their information confidential and the services they use available, despite the difficulty caused by cybercriminals. Because of this, cybergovernance makes all the difference.

What needs to be done is more research on the topic, more cyber governance through law and restrictions, and more vetting on the users of the app. Where we stand right now is with the current tools, we have to help us detect the deepfakes. The more tools that are innovated between now and the future, the better. In conclusion, as we delve deeper into the fakes, we must be sure to trust less and authenticate more. It is also integral that we pay attention, especially in the upcoming years, and check our sources of information. This way we can limit vulnerabilities and have fun while staying safe on the internet!

Works Cited

1]Text - H.R.5586 - 118th Congress (2023-2024): Deepfakes Accountability Act | Congress.Gov | Library of Congress. (n.d.). https://www.congress.gov/bill/118th-congress/house-bill/5586/text (accessed April 30, 2025).

Gil Press. (2024, August 6). Detecting DeepFakes: Fighting AI with AI. Forbes. https://www.forbes.com/sites/gilpress/2024/08/06/detecting-deepfakes-fighting-ai-with-ai/

Harris, Caitlin. “Explore the Top Deepfake Detection Solutions Designed to Identify and Mitigate the Risks Associated with Synthetic Media Manipulation. Explore Features such as Advanced AI Algorithms and Multimedia Analysis.” Expert Insights, 19 Feb. 2024, expertinsights.com/ai-solutions/the-top-deepfake-detection-solutions.

Owda, A. (2025, March 6). Top 10 AI Deepfake Detection Tools to Combat Digital Deception in 2025 - SOCRaDar® Cyber Intelligence Inc. SOCRadar® Cyber Intelligence Inc. https://socradar.io/top-10-ai-deepfake-detection-tools-2025/

“Uncovering Deepfakes: Classroom Guide.” AI for Education, www.aiforeducation.io/ai-resources/uncovering-deepfakes.

University of Virginia. “What the Heck Is a Deepfake? | Information Security at UVA, U.Va.” Security.virginia.edu, University of Virginia, 2023, security.virginia.edu/deepfakes.