Moral Corruption in AI Usage

Many ethical concerns exist; however, not all can be adequately solved as a result of an inability to determine whether a potential measure is most appropriate in some cases. Stephen Gardiner, in his work, “A Perfect Moral Storm” describes an issue of “moral corruption” in which an individual renders themselves incapable of deciding whether their actions are moral, due to complex interactions between a gap between causes and effects, fragmented agencies, and no existing institutions to execute orders. While Gardiner primarily related moral corruption to the issue of climate change in his essay, the same phenomenon can be applied to many modern-day engineering scenarios, including the rise in popularity of deepfake technology, a form of falsified content using an existing image and replacing it with one person’s likeness. As deepfake technology becomes increasingly advanced and accurate, it is also increasingly misused in malicious ways, including impersonation and the creation of harmful explicit photos for intentions such as revenge. Sensity AI, an AI research company, “has consistently found that between 90% and 95% of them are non-consensual” imagery, with a disproportionate number of women represented (Hao, 2021). Despite growing awareness of these downfalls, action taken to regulate the creation of non-consensual explicit deepfakes is complicated due to several challenges including complicated interplays between the general advantages of deepfakes, disadvantages, moral corruption, and motivations for action affected by historical precedent; however, possible solutions still exist.

Increasing moral corruption toward adverse ramifications of explicit deepfakes can be attributed to several factors, including the positive uses of deepfakes or positives of their existence (for society and an individual), negative ramifications of restricting them, and an inability to detect the detrimental outcomes. An important element of the incapability to act toward deepfakes is that they have dual usage; when used properly, they have the potential to benefit society in multiple ways. One of these ways is by allowing healthcare professionals to monitor patient health while protecting patient privacy, an important and favorable use, creating difficulty in what must be encompassed in regulating the products of deepfakes. Another contributing factor when it comes to moral corruption toward the problems associated with deepfakes is that most people cannot tell what content is produced by deepfake technology and what is not, leading to unawareness when consuming the content. Karen Hao of the MIT Technology Review found that 80% of audiences do not even know what a deepfake is, leading many to blindly believe any content they are presented with without any thought otherwise. In addition, Gardiner argues in his work that “many political actors emphasize considerations that make inaction excusable or even desirable” (Gardiner, 2006). This can also be seen in the reasoning that placing limitations on deepfake usage may hinder freedom. The Yale Journal of Law and Technology points out that restrictions on the content may be considered a violation of the First Amendment (of the United States Constitution) due to protections of false statements under the landmark decision *United States v. Alvarez*. Furthermore, it is controversial to restrict what one does in private which can lead to individuals believing that certain rights would be lost by taking action. Moreover, Gardiner describes a “prisoner’s dilemma” in “which one find[s] themselves in a paradoxical position. On the one hand... they understand it would be better for everyone if every agent cooperated; but, on the other hand... they know that they should all choose to defect” (Gardiner, 2006). Instances of this can also be seen with deepfakes, for example, politicians may not want to be impersonated in negative lights; however, the existence of deepfakes creates a way for genuine unfavorable content to be dismissed. An example of this is “Joao Doria, the governor of Sao Paulo in Brazil. In 2018 the married politician claimed a video allegedly showing him at an orgy was a deepfake - and no one has been able to prove conclusively that it wasn't," showing an instance where the existence of explicit deepfakes provided a benefit to a politician rather than a harm (Thomas, 2020). All of these components contribute to moral corruption when it comes to dealing with non-consensual deepfakes.

Non-consensual explicit deepfakes are an intergenerational issue. A large factor in the difficulty in combating deepfakes used malevolently is due to the temporal aspect; for instance, in the case of deepfakes used in non-consensual explicit content, historical precedent and future possibilities create complications that impede the capacity to act. One example of this is the fact that forms of non-consensual manipulated media have existed since long ago, in forms such as images created with Photoshop, and do not have comprehensive regulatory frameworks (Hao, 2021). The lack of historical mandates contributes to an environment where placing new restrictions becomes viewed as a violation of personal freedoms, making it burdensome to implement them in a way that does not encroach upon an individual. Another important matter is that the rapid development of deepfake technology creates a scenario in which lawmakers and technical solutions can no longer keep up (Bond, 2023). Despite attempts to combat this, such as requiring deepfakes to be watermarked, software that may be able to detect deepfakes, "there's not yet a universal standard for identifying real or fake content," which prevents such answers from being able to catch up with the current state of deepfake technology (Bond, 2023). Finally, in the past, people have been morally corrupt to the objectification of other human beings, an issue that disproportionately affects women. This has been ingrained in society for hundreds of years, found in many aspects of it, such as in popular media one consumes, rendering it difficult to solve as it would require a generational shift in belief and additionally, preventing the perpetuation of it. These intergenerational aspects contribute largely to the development of modern deepfakes, and historical context brings about a significant problem in addressing the deepfake problem.

While this issue is undeniably complex with no perfect course of action in its current state, there are a few possible solutions that can relieve the spread of non-consensual explicit deepfakes. One possible resolution may be for social media sites to better monitor uploaded content passing a set duration to their site by reviewing the video using existing deepfake detection tools and either preventing the upload or flagging it, especially on websites such as Reddit, Twitter, or Facebook, rather than reviewing it and removing it later. These platforms bear the responsibility to regulate content uploaded to their sites and prevent the spread of misinformation, as social media sites can place restrictions on media that many governments cannot. This solution may also be time-consuming, however, due to the number of users that create media on some social media sites, therefore it may not be the most perfect solution that exists. In addition to social media sites, individuals must be mindful of the content they consume, and the use of fact-checking and verification of information must be encouraged in the public. It is essential that consumers be critical of the content they encounter to, at the very least, prevent harm done to victims of explicit deepfakes. By combining these efforts, society can begin to tackle this issue and create more ethical digital landscapes.

Non-consensual explicit deepfakes are a growing issue in modern society; however, they can also aid communities, making it difficult to appropriately find ways to limit their use, and increasing the susceptibility of society to developing moral corruption. The temporal aspect of deepfakes also creates a large problem in an institution’s ability to create suitable solutions because of cultural norms and the rapid progress of technology. Despite all of these factors, possible ways to combat their spread and damage can still be done.

References

BLITZ, MARC J. "DEEPFAKES AND OTHER NON-TESTIMONIAL FALSEHOODS: WHEN IS BELIEF MANIPULATION (NOT) FIRST AMENDMENT SPEECH?." *Yale Journal of Law and Technology*, vol. 23, 2020.

BOND, SHANNON. “AI-generated deepfakes are moving fast. Policymakers can't keep up.” *NPR*, 2023, https://www.npr.org/2023/04/27/1172387911/how-can-people-spot-fake-images-created-by-artificial-intelligence. Accessed 6 November 2023.

BOND, SHANNON. “People are trying to claim real vidoes are deepfakes. The courts are not amused.” *NPR*, 2023, https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused. Accessed 6 November 2023.

GARDINER, STEPHEN M. “A Perfect Moral Storm: Climate Change, Intergenerational Ethics and the Problem of Moral Corruption.” *Environmental Values*, vol. 15, no. 3, 2006, pp. 397–413. *JSTOR*, http://www.jstor.org/stable/30302196. Accessed 6 Nov. 2023.

HAO, KAREN. “Deepfake porn is ruining women’s lives. Now the law may finally ban it.” *MIT Technology Review,* 2021, https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/. Accessed 6 November 2023.

THOMAS, DANIEL. “Deepfakes: A threat to democracy or just a bit of fun?” *BBC*, 2020, https://www.bbc.com/news/business-51204954. Accessed 6 November 2023.