Exploring the Ethical Considerations of AI Use for Misinformation and Deepfakes

Gerrett M. Broussard*, Shane T. Crummey, Jr., Kelvin Robinson
*Corresponding Author: gerrett.broussard@my.hamptonu.edu

School of Engineering, Architecture, and Aviation
Hampton University
Hampton, Virginia

***Abstract*— This paper outlines the ethical considerations and consequences of the AI generated images, or deepfakes, through their societal, political, and educational impacts. Primary exploration will be conducted through a Consequentialist perspective, capturing the scale of deepfakes regarding their potential to aid and inhibit humanity. Further exploration will be conducted through the lens of social-contract theory and rights-based ethics to provide a multifaceted exploration of the severity of this issue through its ethical responsibility (ethos) and human impact (pathos). The findings indicate that deepfake technology possesses broader societal risks, despite its benefits in areas such as education, accessibility, and professional training. Research identifies a direct correlation between deepfakes and acceleration in misinformation, fraud, and the erosion of trust in public consciousness. With the accelerating nature of deepfake capabilities, uncertainty continues to grow as pre-established countermeasures continue to be circumvented. From a utilitarian perspective, there is strong support for regulatory framework, development into detection technology, and the encouragement of media literacy to abate this undeniably problematic but irrevocable technology.**

Keywords— Deepfakes, AI-generated media, Ethical implications, Misinformation, Media literacy

I. Introduction

Deepfakes, or artificial media, are hyper-realistic forms of content digitally manipulated to represent humans with inorganic content. This larger category of content is subset by interactive and compositional deepfakes. The former refers to impersonations with realistic multimodal interaction, while the later refers to the amalgamation of other forms of preexisting content to create a new, coherent piece of content [1].

Manipulation in this manner can be performed in a variety of media fields, which are altogether called synthetic media classification. This ranges from video, audio, and images to text-based simulation. Modern deepfake technology is notable for its ability to imitate hyper-realistic characteristics, blurring the line between fabricated and real technology to the untrained eye.

These advances in generative AI pose beneficial use in content, education, and healthcare. Content-creation and media have explored cost-efficient alternatives to previous forms of content generation, such as special effects and advertisements. Deepfakes provide avenues for positive social impact through recontextualized educational and instructional content, engaging previously neglected learning inclinations through animations or multilingual presentations [2]. In healthcare, they are capable of generating lifelike simulations for professional training, granting a wider, safer, and more affordable alternative to traditional training modules [3].

However, these advances also enable malicious activities such as counterfeit, fraud, and political and institutional undermining [4]. This is done through imperceptible identity fraud and social engineering attacks, allowing for the lines between genuine and fabricated content to be blurred more closely than ever before. Most importantly, these deepfakes can have long-term ramifications of uncertainty on the legitimacy of prevalent media [5].

In attempt to contextualize the societal benefit or lack thereof in deepfake-technology, it is important to define the metrics for which such analysis will be conducted. For this paper, Act Utilitarianism will be utilized, which is defined by the Cambridge University Press as the emphasis on producing the greatest overall good for the greatest number [6]. This is a case-by-case analysis, wherein good and right are determined by the overall societal benefit rather than adherence to a rule or principle. This is significant for discussing deepfakes, due to their role as a dual-use technology. With undeniable benefits and detriments and a lack of clarity on legislation defining its usage,

interpreting the usage of deepfakes on a personable level makes it possible to explore the societal impact.

II. Background

A. Deepfake Technology

Deepfakes are defined by Westerlund as AI-generated synthetic media, which leverages multimodal data (video, audio, images, and more) to create lifelike simulations. These are often created using advanced Machine Learning techniques [7]. This is prefaced by Generative Adversarial Network (GAN) technology, which utilizes a two-layered procedure of generation and discrimination. The generator creates synthetic content through the amalgamation of data sets, while the discriminator evaluates the realism of these simulations regarding accuracy. The sophistication of these dual-faceted networks results in the replication of subtle cues such as eye movement, speech iotation, and habitual ticks. In doing so, deepfakes become capable of replicating facial expressions, voice patterns, body language, and more, blurring the line between human and synthetic behavior through multimodal replication [8]. Figure 1 below provides a simplified visual representation of the system [9].

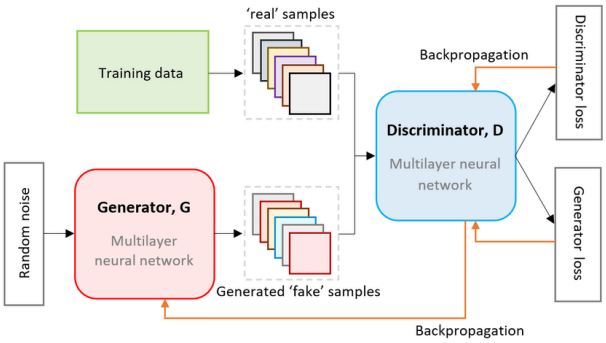

Fig. 1. Demonstration of a GAN procedure.

With the widespread public availability of large-scale data, GAN modules can access vast amounts of information to refine their synthetic outputs [5]. These networks leverage this data to improve the discriminator's ability to distinguish authentic content from its synthetic counterpart, improving the scrutiny of its assessment before generation even begins. As realism improves, this assessment is conflated with human scrutiny. Succeeding deep fakes are made with steps to circumvent traditional methods of manipulation detection, facilitating misuse in various areas [8]. Most pressing among them are misinformation campaigns, political manipulation, identity fraud, and reputational attacks.

B. Emergence in Corporate & Civil Spaces

Recent developments have seen an uptick in deepfake enabled fraud and cybercrime. A common tactic reported by the Bank of America were fraudulent wire transfers, wherein a CEO or CFO's voice would be near-perfectly replicated [4]. Beyond fraud, identity theft has also become a concern, as biometric verification scans such as facial recognition or voice ID are slowly becoming susceptible to deepfake generations [10]. In civil spaces, brand and individual reputations remain at stake due to the ability to fabricate endorsements or statements [1]. Even when disproven, the ramifications often exceed clarification. This was showcased in 2024, during which fake ads were run featuring celebrities such as Tom Hanks, Oprah, and Elon Musk using deepfake technology. This fabricated data lent itself to crypto scams and fake product endorsements, crippling individual trust in these stars in addition to individual trust in the legitimacy of the content they were consuming [11].

C. Stakeholders

Companies facing the highest stakeholder pressure regarding AI ethics are primarily major technology platforms, financial institutions, and large enterprise software developers. Large technology companies such as meta platforms, google, and Microsoft are at the center of this issue because they both develop AI systems and control the platforms where the misinformation spreads. These companies face pressure from governments, users, and advertisers to limit deepfake content while still upholding freedom of expression. Online platforms play a key role in misinformation because of algorithm driven content distribution, which makes them ethically responsible for moderation and detection. [12], [13]. As a result, companies must continue to invest in AI detection tools to reduce harm and maintain user trust.

Financial institutions are also among the most affected stakeholders because deepfakes are being used in fraud and identity theft. Banks and financial tech companies like JPMorgan Chase and PayPal face growing risks from AI generated scams that mimic executives or customers. Reports indicate that deepfakes pose

significant risks to businesses, leading to financial loss and reputational damage [12],[14]. This highlights the ethical responsibility of financial organizations to strengthen verification systems and invest in cybersecurity measures to help protect stakeholders.

Enterprise software developers, including Adobe and OpenAI are also under intense scrutiny because they create tools that can generate deepfakes and false media. These companies must balance innovation with ethical responsibility as their technologies can be used for both good and bad. Concerns about AI misuse and the need for stronger governance. Frameworks continue to grow as deepfake technology becomes more advanced.

Overall, these industries experience the highest stakeholder pressure because they are directly involved in either creating, distributing or being targeted by misinformation produced by AI. The ethical challenges they face, ranging from preventing fraud to maintaining trust, demonstrates that responsibility for addressing deepfakes is shared across numerous sectors, but it falls most heavily on those with the most influence [15].

D. Current Legal Jurisdiction

Current legal jurisdiction surrounding deepfake technology is insufficient, due to both its weak regulation and fragmented enforcement. Efforts to address the societal threat this technology presents are insufficient. This is largely because regulations are limited to state levels. While some states have taken proactive steps, there is no comprehensive federal framework addressing deepfake creation, distribution, and misuse [16].

The lack of consistency makes enforcement near impossible. For example, Texas enacted legislation (Texas SB 751) that criminalizes the creation or distribution of deceptive deepfake videos intended to influence elections or harm individuals. Within them, non-consensual usage of synthetic media can result in misdemeanor or felony charges [17]. California has multiple laws prohibiting the distribution of deceptive media revolving around elections, in addition to enabling civil action lawsuits against non-consensual deepfake generation [18]. These laws, though significant, can only be enforced within these territories, resulting in enforcement gaps. The language of these laws presents additional challenges, as the act of proving ill intent or harm remains ambiguous and difficult to consistently apply.

Beyond nation-wide enforcement, the globalization of deep-fake distribution poses additional challenges. Enforcement of deepfake regulation necessitates international cooperation, yet global standards remain underdeveloped. As of 2026, there is still no unified global legal standard specifically targeting deepfake misuse [19]. As technology continues to evolve, circumventing many of these regulations and methods of detecting its generation, legislative efforts fail to keep up. This results in many laws being enacted only after significant harm has already occurred [17].

III. Utilitarian Analysis

A. Observable Benefits

Deepfakes allow for highly engaging, personalized content. This is essential in communities who, either due to resources or location, would otherwise be deprived of these opportunities [3]. Martínez et al. identify the possibility of recreating historical figures or modern scientists to add interactivity to lessons in a way standard videos cannot. This is significant as real-time interaction plays a critical role in student engagement, and by extension retention, and conceptual understanding. Studies have shown that learning engagement improves up to 30-40% with avatar-based instruction This makes deepfake-driven education a vital alternative to passive video instruction, especially for younger audiences [2].

Older students can also benefit from deepfake education, particularly through training simulations. In healthcare education, simulated deepfake-based training environments allow repeated exposure to rare or high-risk scenarios (e.g., cardiac arrest complications or surgical anomalies), which are otherwise inaccessible in traditional clinical rotations due to ethical and logistical constraints [3]. Operations on living patients require preliminary expertise and limited availability. Mannequins and lab sessions can cost thousands of dollars and have similar issues regarding availability and accessibility. Simulations can offer repeatable, safe methods of practice for students and professionals. They also allow for a wider coverage of practice methods, as they can create abnormal scenarios typical practicing methods cannot.

Martinez highlights that nursing students reported increased confidence and preparedness after performing simulated patient interactions, particularly through synthetic media, especially ones regarding communication heavy scenarios such as patient counseling and emergency responses. Without deepfake content generation, these scenarios would be difficult to replicate [21].

Beyond education, deepfakes can also aid in accessibility and inclusion. Through voice and video modification multilingual media generation becomes possible. This allows for a single instructional video to be adapted across dozens of languages while maintaining natural visual coherence, which by extension expands global accessibility [3]. Duolingo has already begun to adapt this concept through its use of AI-driven voice synthesis in language-learning content. Deepfake usage can go beyond language barriers into physical ones, through the transformation of spoken lectures into sign-language avatars or emphasizing lip intonations for the hearing impaired. This makes it possible to instruct those with hearing or visual impairments or other disabilities [6].

In an economic sense, they provide corporations and creators with tools to reduce onboarding and marketing costs. From a production standpoint, these tools can reduce post-processing costs by 50-80% through the reduction of reshoots and localization [1]. In place of human spokespeople, which are limited by studios, travel, and production teams, AI-generated voices enable scalable marketing campaigns with a greater amount of control and long-term retention [10]. In an age of growing globalization, this ease of access is essential in reaching foreign markets. In place of post-processing and virtual effects, deep fakes can be used to create highly realistic content with a reduced level of expertise and technical demands [1].

From a utilitarian perspective, these applications demonstrate the ability of synthetic media to expand access to education by increasing accessibility and reducing cost. They also provide a new avenue for businesses in the form of consistent, scalable media.

B. Observable consequences

Deepfakes result in an erosion of trust in the media. Recent studies indicate that nearly 48% of individuals have begun the authenticity of content they encounter online [12]. Through the trust-erosion synthetic media presents, by making it difficult to differentiate real from fake content, skepticism of legitimate news increases [8]. This reduction of trust has a direct harm on journalism, legal evidence, and even personal communication.

For instance, after news of Nicolás Maduro's capture in early 2026, social media platforms were flooded with AI-generated deepfakes showing the Venezuelan leader in fabricated situations, such as in jail with the rapper Sean "Diddy" Combs, drawing millions of views and spreading confusion over what was real and what was machine-generated [12]. Events like these, even after clarification occurs, decrease trust and create uncertainty in news consumers and reporters. When time is of the essence, yet information continues to be fabricated, balancing immediate commentary with accuracy becomes increasingly difficult.

Beyond erosion of trust, deepfakes also have sinister implications in their usage for fraud, impersonation, and security risks—as previously mentioned. They also pose psychological and social harm. In one documented case, a multinational company lost more than $500,000 after employees were deceived by fabricated videos of their company leadership. Synthetic content makes the facilitation of fraud easier through scalability, ignoring the once prevalent barrier of human operations. Even worse, these scams can take on a more intimate stature. 77% of individuals targeted by scams involving the impersonation of family members or trusted figures have reported financial losses, highlighting not only the quality of deepfake content, but its ongoing misuse.

Victims of explicit deepfake generation or libel experience mental and social consequences, some of which result in irrevocable damage. California students were observed creating and sharing non-consensual deepfakes as a form of harassment, leading to social ostracism, blackmail, and mental health struggles with no true recourse [13]. Even if clearly false, the nature of the fabricated allegations can damage careers and personal relationships. The social stigma and doubt often persist, and individuals

can feel alienated over matters beyond their control [8]. More than anything, the simple existence of deepfake technology increases anxiety about media authenticity [14]. This results in general doubt over real images and videos, and contributes to social distrust.

From a utilitarian perspective, deepfakes present a collective reduction in societal wellbeing. As trust diminishes, the efficiency of communication, reliability of institutions, and stability of social relationships are all compromised. Beyond the numeric, quantifiable number of victims of synthetic media, the social impact these deepfakes present is just as if not more concerning. The cumulative effect is a systemic weakening of informal infrastructure, one that requires more time, effort, and resources to be pooled just to reach a once-normal level of credibility.

C. Cross-Analysis

An Act-Utilitarianism perspective affirms that the greatest perspective on deepfakes weighs their inherent potential and historic use for good with their risks and propensity for malpractice. Within this framework, it is undeniable that deepfakes demonstrably enhance education, accessibility, and professional training. This is done by improving engagement, reducing costs, and expanding costs with regard to globalization and accessibility concerns.

Despite these clear benefits, there are several concerns worth identifying. Empirical data links deepfake developments with a significant erosion of trust and systemic decline in confidence within digital information ecosystems. Deepfake-enabled fraud has surged dramatically, resulting in substantial financial losses and social ramifications for both individuals and institutions. These issues extend beyond quantifiable damages into psychological harm and the destabilization of truth itself. A difficult to quantify but impossible to ignore is that of the ‘liar’s dividend’, wherein authentic evidence can be dismissed as fabricated. Its potent threat necessitates further substantiation within any claims made with digital media, creating an exhausting sense of uncertainty regarding one’s own work.

Weighing the benefits to detriments, there is a clear distinction in scope. While the benefits of deepfakes are centralized around education and the commercialization of businesses, the ramifications of synthetic media can be felt in every aspect of the digital landscape. The greater and more quantifiably significant reach of deepfake malpractice indicates that its existence produces more harm than good. From that perspective, an act-utilitarianist perspective assesses deepfake technology as something to carefully regulate with regard to its dangers, rather than something to embrace due to its benefits.

IV. Avenues for Potential Solutions

Effective mitigation of deepfake risks requires a multi-pronged approach. This can be achieved by focusing on technological detection, regulatory framework, media literacy, and corporate vigilance. The same research that identifies the risks of synthetic media also identifies strategies that can mitigate their impact. Implementing these solutions would make it possible to embrace the undeniable benefits of deepfake technology on education, accessibility, and more.

Technological detection serves as the first line of defense. The same GAN network discrimination detection can be utilized to scan media for inconsistencies, providing an immediate way to identify and flag certain media for human scrutiny [9]. Such systems would inhibit prevalent fraud, identity-theft, and large-scale scams which utilize volume and naivete to overwhelm traditional detection systems. Digital watermarks and content verification protocols provide additional safeguards by ensuring that media can be traced and authenticated, allowing fabricated content to be efficiently removed or traced back to its source [8]. These two measures would inhibit malpractitioners, and while the threats of deepfakes wouldn’t be removed entirely, they would serve to exacerbate the efforts required to perform sustained attacks with them.

Regulatory and legal measures serve as a deterrent against malicious use. Stronger cybersecurity and anti-fraud laws can directly criminalize the intentional creation and distribution of harmful deepfakes [10]. Because many deepfake-enabled crimes occur across borders, international collaboration is essential to

establish standards, share intelligence, and prosecute perpetrators effectively [20].

The final and most important step is enhancing media literacy and public awareness. The aforementioned efforts are meaningless if individuals remain incapable of identifying synthetic content. Education campaigns, clear labeling of AI-generated media, and training would serve as the first steps in developing scrutiny. By weaponizing public discernment against deepfakes, the effectiveness of manipulative content can be reduced [20]. Subtle prompts or flags and resources on social media sites can enhance public vigilance.

Together, these strategies aid in mitigating harms such as fraud, misinformation, and social distrust. Through them, they can preserve the legitimate, beneficial applications of deepfakes while minimizing the negative consequences that threaten overall well-being. While risk is unavoidable, proper global policy and education can aid in ensuring that the benefits of synthetic content outweigh these vices.

V. Conclusion

Deepfake technology refers to synthetic media created by GAN machine learning models, harnessing generation, and evaluation to replicate organic media. This technology represents a dual-faceted innovation, serving to both enable meaningful societal advancement while posing significant utilitarian detriments. When approached through an act-utilitarian lens, the potential uses of deepfake technology within education and business are offset by their potential to erode trust across digital media landscapes through misinformation and fraud. Rather than eliminating the technology and its potential benefits, a more realistic approach would be to leverage resources to inhibit malpractice. By improving detection technologies, strengthening regulatory measures, and fostering widespread media literacy, it will be possible for synthetic and legitimate media to coexist across the digital landscape. Ultimately, the goal is to preserve truth, security, and trust within this critical period of technological advancement.

Acknowledgment

Thank you to Hampton University's School of Engineering, Architecture, and Aviation and research funding from the University of Virginia for conference registration.

References

[1] S. L. Burton and D. P. Harvie, "Deepfakes: Unmasking the technological, societal, and ethical dimensions," RAIS Journal for Social Sciences, vol. 9, no. 2, 2025. [Online].

[2] H. Sharif, A. Atif, and A. A. Nagra, "Deepfake-style AI tutors in higher education: A mixed-methods review and governance framework for sustainable digital education," Sustainability, vol. 17, no. 21, 2025. [Online].

[3] O. Navarro Martínez, D. Fernández-García, N. Cuartero Monteagudo, and O. Forero-Rincón, "Possible health benefits and risks of deepfake videos: A qualitative study in nursing students," Nursing Reports, vol. 14, no. 4, 2024. [Online].

[4] Bank of America, "Deepfakes: Business risks and mitigation strategies," Bank of America Cyber Security Journal, 2023. [Online].

[5] Keepnet Labs, "Deepfake statistics and trends: Key data and insights," 2025. [Online].

[6] J. J. C. Smart and B. Williams, Utilitarianism: For and Against. Cambridge, U.K.: Cambridge University Press, 1973.

[7] T. Westerlund, "The Emergence of Deepfake Technology: A Review," Technology Innovation Management Review, vol. 9, no. 11, pp. 39–52, 2019.

[8] S. Alanazi et al., "Unmasking deepfakes: A multidisciplinary examination of social impacts and regulatory responses," Human-Intelligent Systems Integration, vol. 7, pp. 131–153, 2025, doi: 10.1007/s42454-025-00060-4.

[9] A. Mirza and S. Osindero, "Example of GAN architecture," in *Unsupervised and Transfer Learning Challenge: A Deep Learning Approach*, ResearchGate, online.

[10] G. Singh and R. Patel, "Navigating the Deepfake Threat: Cybersecurity, Ethical Implications, and Legal Challenges," Journal of Innovative Engineering Research, vol. 5, no. 2, pp. 1–10, 2024.

[11] B. Colman, "The Impact of Deepfakes on Brand and Reputation," Reality Defender, Aug. 07, 2024. [Online].

[12] IBM, "What are deepfakes?" 2024. [Online]

[13] National Institute of Standards and Technology, "Evaluating deepfake detection and AI risks" 2024. [Online].

[14] Brookings Institution, "How deepfakes undermind democracy and trust," 2023. [Online].

[15] M. Bowen, "iProov study reveals deepfakes shatter online confidence," Intelligent CISO, Mar. 10, 2026. [Online].

[16] R. Chesney and D. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," California Law Review, vol. 107, no. 6, pp. 1753–1820, 2019. [Online].

[17] Texas Legislature, "SB 751: Relating to the creation and distribution of a deep fake video with intent to injure a candidate or influence an election," 2019. [Online].

[18] California Legislature, "AB 730 and AB 602: Deepfake laws on election interference and non-consensual pornography," 2019. [Online].

[19] European Commission, "Artificial Intelligence Act," 2024. [Online].

[20] "AI deepfakes of Nicolás Maduro flood social media — depict Venezuelan dictator in jail with Diddy, among other vids," New York Post, Jan. 6, 2026.

[21] C. Vaccari and A. Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," Social Media + Society, vol. 6, no. 1, Feb. 2020