GLOBSEC
IDEAS SHAPING THE WORLD

Regulating Deepfakes: Global Approaches to Combatting AI-Driven Manipulation

Policy Paper

Jana Kazaz, Centre for Democracy & Resilience

1. Introduction

Deepfake technology, a product of artificial intelligence (AI), has rapidly emerged as a powerful tool for creating hyper-realistic, yet manipulated, audio, video, and image content. These digital forgeries are becoming increasingly sophisticated, making it difficult for the average viewer to distinguish between authentic and fabricated media. While deepfake technology has potential **benefits** in entertainment and creative fields, its misuse poses significant risks to society, ranging from personal identity theft to the undermining of democratic processes.

Deepfakes have rapidly become a global concern. In recent years, the amount of deepfake content online has **surged**, with reports showing that the number of deepfake videos grew by 900% between 2019 and 2020. By 2026, some researchers have **predicted** that as much as 90% of online content may be synthetically generated. Given the global nature of the internet and the ease with which deepfake content can spread, governments worldwide have started considering or enacting legislation to curb the misuse of this technology.

Currently, regulatory approaches to deepfake technology **vary** significantly across the world. Some jurisdictions, such as the European Union or China, have adopted overarching AI regulations. In the United States, the approach is more fragmented; while no federal law addresses deepfakes or AI in general, several states have enacted their own legislation, focusing on specific applications like election security or explicit content. Other countries, such as Taiwan or the UK, have opted to amend criminal laws in specific areas of use of deepfakes, to combat the misuse of deepfake technology in fraud cases (Taiwan) or non-consensual intimate images (the UK). Finally, some nations rely on existing legal protections for image rights and privacy without specific new regulation. Italy, for instance, uses its constitutional and civil codes to address deepfake misuse, leveraging established legal frameworks to protect against identity infringement and reputational harm.

This policy paper explores and compares these diverse regulatory approaches, drawing on examples from the EU, the United States, China, Taiwan, the UK, and Italy. It highlights the strengths and gaps in existing frameworks and offers recommendations to policymakers on crafting regulations that deter misuse while fostering innovation and safeguarding fundamental rights.

2. What is Deepfake Technology and Why it is Dangerous

Deepfake technology **refers** to AI-manipulated media, such as videos, photos, or audio recordings, that appear real but have been altered to depict people saying or doing things they never actually did. The underlying technology can replace faces, manipulate facial expressions, synthesise speech, and more. Deepfakes rely on advanced AI techniques, notably Generative Adversarial Networks (GANs) and autoencoders. GANs consist of two competing neural networks—one that generates fake content and another that attempts to detect it. This constant **competition** results in highly realistic but fabricated content that can be hard to detect.

Deepfakes have already been exploited for malicious purposes, with significant real-world impacts in areas like misinformation, identity theft, and fraud. In the political arena, deepfakes have been used to create false media depicting political candidates saying or doing things they never did, which can spread disinformation during elections and erode public trust in democratic processes. For example, in India's 2024 elections, AI-generated deepfakes were actively **used** by political campaigns to clone candidates' voices and create holographic avatars, to further their reach to voters. At the same time, however, there was a lot of AI-driven misinformation **used**, particularly on platforms like WhatsApp and YouTube, where low-quality deepfakes spread quickly with a potential to incite violence and riots. Similarly, in Moldova, deepfake videos falsely **showed** the country's president endorsing a pro-Putin party during local elections, while in Taiwan, Chinese state-backed actors used AI-generated audio to falsely claim that a prominent politician had **endorsed** a rival candidate. These are just few examples out of many, as by the end of the super-electoral year 2024, it is rare to find a country with elections untouched by the misuse of deepfake technology, which poses significant challenges to electoral integrity worldwide.

Deepfakes also **pose** significant personal risks, particularly in the creation of non-consensual pornography, where women are disproportionately victimised. These fabricated materials exploit individuals' likenesses, violating privacy and creating lasting social harm. Beyond individual exploitation, deepfakes have **become** a major threat to financial systems. Cybercriminals have used deepfake audio and video to trick employees into authorising fraudulent transactions, highlighting the technology's potential to exacerbate corporate fraud.

The growing threat of deepfakes underscores the urgent need for both technological and legal solutions to mitigate their harmful effects. Without adequate detection tools and regulatory frameworks, deepfakes could significantly erode trust in digital media, political systems, and even interpersonal communication. For example, as seen during India's 2024 elections, the surge in AI-generated content, coupled with the lack of comprehensive regulation, has allowed misinformation to spread unchecked, posing serious risks to the integrity of democratic processes.

3. Comparative Regulatory Landscape

• United States: Regulatory Patchwork

◗ Federal Level:

The United States lacks comprehensive federal legislation specifically addressing deepfakes. Although several proposals have been introduced, such as the **Deepfakes Accountability Act** and the **Protect Elections from Deceptive AI Act**, none have been enacted into law yet. The former aims to protect national security by imposing transparency requirements on AI-generated content and providing legal recourse for victims of harmful deepfakes. The latter proposal seeks to curb the use of AI-generated deceptive materials in political campaigns by amending the Federal Election Campaign Act of 1971 ("FECA") to prohibit the distribution of misleading

audio, images, or videos involving federal candidates in political or issue-based ads.

The Protect Elections from Deceptive AI Act (Act) imposes a general ban on individuals, political committees, and other entities knowingly sharing materially deceptive AI-generated content. However, to ensure it does not infringe on First Amendment rights, the act includes key exemptions for content that are clearly labelled as parody or satire, as well as for legitimate news broadcasts. Additionally, radio and television stations that broadcast such content with proper disclosures are exempt from this prohibition.

The proposed penalties under the Act aim to deter the use of AI-manipulated media in elections, with violators facing significant fines based on the extent of the deception. Federal candidates whose likeness or voice has been misrepresented can seek legal recourse in federal courts, including removal of the content and compensation for damages.

In the realm of personal abuse, such as non-consensual pornography, the absence of comprehensive federal legislation leaves victims reliant on existing laws addressing harassment, privacy, or defamation. Proposals like the Deepfakes Accountability Act include provisions aimed at addressing this issue, but these remain stalled at the federal level. Meanwhile, the Federal Trade Commission (FTC) has recently proposed new rules targeting AI-driven impersonation, which could apply to deepfakes in fraudulent or abusive contexts. However, the absence of specific deepfake-focused federal laws results in inconsistent protections across the country.

In cases of fraud, federal laws addressing impersonation and identity theft are sometimes applied to deepfake misuse, but their scope is limited. The FTC's push to expand protections against AI impersonation highlights the growing recognition of this gap, though enforcement mechanisms remain underdeveloped.

State Level:

At the state level, there has been a growing focus on regulating deepfakes, especially in the context of political ads and election integrity. As of September 2024, 23 states[1] have **passed** legislation addressing deepfakes, all of them with bipartisan support, with 17 states enacting bills or amendments this year alone.

State legislation addressing non-consensual pornography includes pioneering efforts such as Virginia's law, which criminalises the distribution of deepfake pornography as a Class 1 misdemeanor, punishable by up to one year in jail and a $2,500 fine. California has also enacted laws that allow victims to sue for damages if their likeness is used in deepfake pornography without consent, providing some recourse for individuals affected by this abuse. However, these protections remain patchy, and victims in states without similar laws often face significant legal challenges.

In cases of fraud, some states, such as Texas, have enacted legislation that explicitly criminalises the use of deepfake technology to deceive or defraud others. These laws serve as important steps but highlight the fragmented nature of state-level approaches, which result in varying levels of protection for victims depending on their location.

State laws addressing election integrity typically fall into two **categories**: disclosure requirements and temporary bans on deepfake content in the run-up to elections. Several states mandate clear labelling of AI-generated content in political ads. For instance, Michigan **mandates** continuous disclaimers within 90 days of an election, with violators facing misdemeanour charges, fines of up to $1,000, or up to 3 months in jail. Wisconsin **imposes** similar penalties, while New Mexico **combines** disclosure requirements with public awareness campaigns to educate voters on deepfake risks. Arizona **focuses** on a narrower approach, targeting digital impersonation of candidates or officials and allowing civil suits against offenders.

Some states introduced temporary bans on deepfake content, which prohibit the distribution of AI-

1 AL, AZ, CA, CO, FL, HI, ID, IN, MA, MI, MN, MS, NH, NM, NY, OR, TX, UT, WA, WI

generated political content without proper disclosure during specific timeframes leading up to elections. Hawaii, **California**, and **Michigan** have enacted such restrictions. For example, Hawaii's law bans materially deceptive content during election years from February to Election Day unless accompanied by a clear disclaimer. Critics **argue** that such restrictions should apply year-round, not just during election cycles, as misleading depictions of real individuals, whether politicians or not, pose a constant threat.

• China: Comprehensive Control

China has **taken** a comprehensive approach to regulating deepfake content through the *Provisions on the Administration of Deep Synthesis of Internet Information Services*, which were **adopted** in January 2024. The goal of this legislation is to preserve social stability, making it distinct from the narrower deepfake regulations seen in the United States and Europe. The *Provisions* **cover** various forms of AI-generated content, including text, images, audio, and video, addressing both the technological and social challenges posed by deepfake technology.

One of the key requirements under the new law is the mandatory labelling of all AI-generated content. All AI-manipulated media must be clearly marked with a watermark or textual indication, ensuring viewers are aware that the content has been altered. Additionally, the production of deepfakes without user consent is strictly prohibited, protecting individuals from unauthorised use of their likeness or personal data. The legislation places obligations on both platform providers and end-users. Platform providers offering content generation services must take responsibility for the ethical use of AI, which includes evaluating and verifying the algorithms used, authenticating users to track content creators, and implementing feedback mechanisms for consumers. Furthermore, platforms are required to avoid processing personal information, in alignment with China's broader data protection policies.

This regulation follows a broader trend of Chinese authorities relying on technology companies to contribute to social and political stability. Chinese tech firms, such as Alibaba and Tencent, have been **compelled** to align with the government's 'common prosperity' objectives, which emphasise economic and social stability. This approach places immense pressure on platforms to cooperate with government mandates, as non-compliance can lead to severe consequences, such as fines, suspension of service or even a criminal liability.

China's regulation of deepfakes **goes** beyond addressing technical issues, reflecting its broader goal of controlling the internet to ensure national security. Unlike the United States, where free speech protections limit the ability to impose radical deepfake regulations, China's government moves swiftly to manage emerging technologies. This expansive regulation underscores China's reliance on tight control over digital content, as the government sees any uncontrolled use of AI technologies as a direct threat to the stability of the regime.

Despite these efforts, the absence of unified federal legislation leaves the regulatory landscape fragmented, with gaps in protections against personal abuse, fraud, and election interference. A comprehensive federal approach would be critical in addressing these challenges systematically and consistently across the United States.

• The European Union: Transparency, and Risk Management Under the AI Act

The EU has introduced the Artificial Intelligence Act (AI Act), a framework designed to foster trustworthy AI while addressing risks associated with AI-generated content, including deepfakes. Rather than imposing an outright ban, the AI Act mandates transparency for those creating or using deepfakes, requiring them to disclose the artificial origin of the content and provide information on the techniques used.

The Act categorises AI systems into four risk levels: unacceptable, high, limited, and minimal. Deepfakes are typically classified in the "limited risk", because of their potential to deceive individuals who may struggle to recognise them as AI-generated. The transparency obligations, outlined in Article 50 (3), require that AI-generated content is clearly labelled, ensuring individuals are aware when interacting with

such media. However, deepfakes that could be used to influence elections or political processes might be **classified** as "high-risk," as stated in Annex III, particularly when they could disrupt the democratic process. However, the exact decision-making criteria for this classification are not specified, yet.

Critics **argue** that disclosure requirements alone may be insufficient to address the risks posed by deepfakes, especially considering their transnational nature and rapid proliferation. While the AI Act is still in its early stages of implementation, the effectiveness of its enforcement mechanisms remains to be seen. The establishment of the EU AI Office in February 2024 is a step toward developing enforceable rules that address these challenges. Experts **emphasise** the need for a clear liability framework within the EU, supported by robust enforcement measures, including criminal liability for individuals who create or distribute deepfakes to manipulate, harass, or defraud others.

Approaches in other countries

• Taiwan's Approach: Criminalizing Deepfake Fraud

In addition to the regulatory approaches to deepfakes discussed above, some countries have chosen to address the problem in a more targeted way, criminalising the most harmful forms of deepfakes. Taiwan, for instance, has taken a more targeted approach to regulating deepfake technology, particularly focusing on its use in fraudulent activities. On May 16, 2023, Taiwan's Legislature **passed** amendments to the criminal law that specifically address the use of deepfakes in fraud, to target the creation and dissemination of computer-generated images, voices, and magnetic records used to deceive individuals. These amendments significantly **increase** the penalties for those found guilty of using AI-generated media to deceive others, with offenders now facing up to seven years in prison and fines up to NT$1 million (approximately US$32,462).

• The UK's Response: Protecting Against Non-Consensual Deepfakes

UK has prioritised addressing the risks of AI-generated sexually explicit images with the **Online Safety Act 2023**. To further strengthen this regulation, in 2024 the UK government **announced** plans to elevate the sharing of intimate images without consent, including deepfakes, to the status of a "priority offence," placing it on par with other serious online crimes such as the sale of weapons and drugs. Under the amendment, online platforms will be required to take faster and more robust action to prevent, detect, and remove non-consensual intimate imagery. While the legislation **places** significant responsibility on platforms to implement stringent measures to combat these violations, it also holds individual users accountable. Those who create or share non-consensual intimate images can face prosecution under existing laws addressing image-based abuse or harassment.

• Italy: Addressing Deepfakes Through Existing Laws

In Italy, deepfakes are **regulated** under existing laws related to personal rights, image protection, and privacy. Article 2 of the **Italian Constitution** protects personality rights, including the right to control one's image, while Article 10 of the **Italian Civil Code** prohibits the unauthorised use of a person's likeness. Combined with Articles 96 and 97 of the **Italian Copyright Law**, these provisions allow individuals to claim compensation if their image is used without consent, particularly when it harms their honour or reputation. A notable precedent illustrating the application of these laws **involves** Prime Minister Giorgia Meloni, who pursued legal action against individuals accused of creating and distributing deepfake pornographic videos featuring her likeness. In this instance, Meloni sought a symbolic compensation of €100,000, despite the indictment revealing that the videos had been viewed millions of times worldwide over the course of several months online.

While democratic countries universally protect personal rights, adapting these protections to the challenges posed by AI-generated content, such as deepfakes, requires specific procedures and clear guidelines. Existing laws theoretically provide a basis for addressing deepfakes, however, as demonstrated by the case of Prime Minister Meloni, their application is often not fast enough to mitigate the harm effectively.

4. Recommendations for Regulators

Prioritise Criminalisation of Harmful Deepfakes: The criminalisation of harmful deepfakes is essential and should cover all cases where deepfakes are produced with the intent to manipulate, harass or defraud others. While existing criminal laws may address fraud or harm to personal rights, it is crucial to have specific legislation dedicated to deepfakes given the new technology and its rapid advancement. Laws should also include clear definitions of deepfakes as opposed to other forms such as AI- altered content or cheapfakes. Responsibility should lie with those who create or spread such deepfakes with malicious intent. Types of deepfakes that should be regulated:

- **Non-Consensual Exploitation and Defamation:** Deepfakes that use someone's likeness or voice without consent and with the intent to harm, defame, or exploit should be criminalised. Such misuse is a violation of personal rights and can cause lasting reputational damage. For example, the UK's Online Safety Act 2023 addresses this by requiring platforms to remove non-consensual intimate images, including deepfakes, to protect users from exploitation.
- **Fraud and Identity Theft:** Deepfakes used in fraud or identity theft should be targeted with specific criminal penalties. Taiwan's recent amendments that penalise deepfake fraud provide a model for addressing these risks within criminal law.
- **Disrupting Democratic Processes:** Deepfakes aimed at swaying public opinion, interfering with democratic institutions, or undermining electoral integrity — such as using deepfakes to incite violence or deter voters — should face strict legal penalties.
- **Public Health and Security Threats:** Deepfakes that spread false information on public health or create security risks—such as fabricated emergency alerts or falsified health data.

Introduce Clear Procedural Guidelines and Effective Penalties: At the same time, effective deepfake regulation requires establishing clear procedural guidelines to navigate this new landscape. Stakeholders, from law enforcement to judicial authorities, as well as potential victims, need straightforward instructions on how to respond if they encounter a harmful deepfake. This includes guidance on reporting procedures for individuals whose likenesses are manipulated for explicit content and for political candidates targeted with disinformation. Procedural rules should make it clear where to report such incidents and outline the steps authorities will take to investigate and mitigate the harm. Additionally, these frameworks should include penalties with a sufficient deterrent effect to discourage misuse, ensuring that those who create or distribute harmful deepfakes face meaningful consequences.

Ensure Comprehensive Transparency Measures: Not all deepfakes should be outlawed, however, transparency measures are crucial to balancing their legitimate use for creativity or entertainment with safeguards against harm and misuse. Clear guidelines should ensure that AI-generated content is identifiable without stifling innovation. Transparency regulations should include mandatory year-round labelling for political campaigns, building on examples from U.S. states like California and Texas, which require disclaimers for AI-generated political ads. These regulations should apply throughout the entire election cycle, not just during a limited pre-election window. Extending regulations year-round would better protect democratic processes and prevent manipulation at any time.

Promote Public Awareness and Education: Governments should launch public awareness campaigns to educate individuals on the risks associated with deepfakes and how to recognise AI-generated content. This can be done by providing resources to help citizens identify manipulated media, as seen in New Mexico's initiative to raise awareness about deepfake threats. Educating the public is crucial to reducing the spread of misinformation and mitigating the social harms of deepfake technology.

5. Conclusion

Deepfake technology represents both a remarkable achievement in artificial intelligence and a significant societal challenge. While its creative and educational potential should not be overlooked, the harm caused by its misuse—ranging from personal exploitation to large-scale disinformation—underscores the urgent need for tailored and effective regulation. This paper has illustrated the diversity of regulatory approaches across jurisdictions, from comprehensive frameworks to more targeted measures, reflecting the unique priorities and legal systems of each region.

As the technology continues to evolve, a proactive and collaborative global approach is necessary to address its risks. Policymakers must prioritise criminalising malicious uses of deepfakes, establish clear procedural guidelines for enforcement, and promote transparency measures that ensure AI-generated content is identifiable. At the same time, fostering public awareness and encouraging cross-sectoral innovation in detection technologies will be essential to staying ahead of emerging threats, in which all – state, private and non-governmental sectors must play a role.

Ultimately, the success of deepfake regulation will depend on striking a balance—one that preserves individual rights, protects societal integrity, and fosters technological progress. By taking decisive action today, governments and stakeholders can mitigate the risks of deepfake misuse and build a digital ecosystem grounded in trust and accountability.