The Impact of AI Deepfakes on THE INTERNATIONAL Right to Human Dignity and Identity

Abstract

This study examines the growing impact of deepfakes and non-consensual synthetic media on human dignity and personal identity, while assessing the adequacy of existing legal and governance frameworks. It argues that although generative artificial intelligence offers significant creative and technological opportunities, it also creates serious risks such as identity manipulation, misuse of image and voice, non-consensual intimate content, reputational harm, identity theft, and the spread of misinformation. The findings suggest that current legal frameworks, particularly at the international and cross-border levels, are insufficient to address the speed, complexity, and multilayered nature of these harms. The article further emphasizes that tools such as transparency measures, labeling, and watermarking, while useful, are not sufficient on their own to remedy the psychological, social, and reputational damage caused by harmful deepfakes. Ultimately, the study calls for a comprehensive, rights-based, and victim-centered approach in which human dignity serves as the guiding principle for the regulation of artificial intelligence and synthetic media.

Keywords: Artificial Intelligence**,** Deepfakes**,** Human Dignity**,** Personal Identity

Table of Contents

Chapter 1: Introduction

1.1. Background and Context

Generative AI has emerged as a transformative technology capable of producing text, images, audio and video, and has quickly gone from a specialized technology to something that's widely accessible. Generative AI differs from the typical AI approaches that primarily classify or predict move action to create new content based on recognising patterns from vast amounts of data. The introduction of synthetic media has enabled more realistic, scalable, and user-friendly solutions thanks to the advancements in diffusion models, large language models, and multimodal systems (Taeihagh, 2025; Ul Islam & Fang, 2026).

AI-generated or AI manipulated text, images, speech, music or video is considered synthetic media. One such prominent example is deepfakes, which have the potential to accurately portray a recognizable individual saying or doing something without their knowledge. Their applications have become more persuasive because of the enhancement of technologies for facial reenactment, voice cloning, lip-syncing and text to video systems (Ul Islam & Fang, 2026).

Media from synthetic material also has advantages in creativity, education, accessibility, translation and entertainment. For that reason, it is crucial that in addition to discussing harmful aspects of synthetic media, innovations with a societal value be discussed and preserved by governance (Taeihagh, 2025; 'Regulatory Challenges in Synthetic Media Governance,' 2024).

Concurrently, these facilitate the manipulation and abuse of identities, intimate imagery without consent, non-consensual intimate imagery, misinformation, and reputational damage which can be committed through impersonation and fraud. Governance should focus on the medium in addition to training data, platforms, and existing options to remedy the situation, as the production of synthetic media is particularly inexpensive, and it can be highly disseminated as it can be made very quickly (Rayun et al., 2026; Taeihagh, 2025).

This change is to be seen as a social and technical change. A "synthetic" media poses questions about what is credible and reliable in the eyes of the public, questions for individual rights, institutional trust and digital communication (Leibowicz, 2026; Rayun, et al., 2026).

1.2. Statement of the Problem

Absolutely, this is a serious human-rights issue: a deepfake has the potential to depict a person's face, voice, and identity — in a face that belongs to someone else and in situations completely unrelated to what they would ever want. Thus, the identity of a person is made an object to be manipulated, misled or exploited for profit.

Such appropriation can be a violation of the 'dignity' of a person, as they become subject to digital appropriation. According to Goodyear (2025), the harms of a deepfake are dignitary harms such as loss of identity control, reputational harm, shame and social exclusion. If false information is revealed, damage can continue since content can be copied and redistributed.

Thus, a dignity-centred approach is crucial. Scharffs and Brown propose a framework for evaluating the use of AI through the lenses of dignity-enhancing, dignity-neutral, or dignity-degrading. As part of synthetic media, non-consensual impersonation, sexualised fabrication and humiliating manipulation can be harmful even when the viewer is aware that the media is synthetic.

Another aspect of the implications of the deepfake is the realm of personal identity, which encompasses the concepts of privacy, autonomy, reputation, informational self-determination and image rights. These risks vary by geographic region or population. Vulnerable populations such as the female and child populations, queer populations, those who are public-facing, and other groups are disproportionately at risk, particularly with the advent of Pornographic deepfakes (Goodyear, 2025; Saxena & Pathak, 2025).

There is a lack of coherence in the existing legal remedies. Standard privacy and defamation laws, data protection laws, harassment laws, and intellectual property laws might only partially raise the issue, but they will not fully capture the issue of identity appropriation and the harm to the person. The tools of transparency (so-called) like labels, or watermarks do not recreate dignity, autonomy or a sense of control over one's identity (Goodyear, 2025; Leibowicz, 2026).

1.3. Research Questions and Objectives

This study asks:

To what extent do existing international human-rights and AI-governance frameworks adequately protect human dignity and personal identity against harms caused by deepfakes and other forms of non-consensual synthetic media?

It examines:

- how dignity, identity, privacy, autonomy, and reputation should be understood in relation to AI-generated representations;
- what harms arise from deepfakes and biometric cloning;
- how existing human-rights and AI-governance frameworks respond;
- whether transparency measures are sufficient; and
- what legal and policy reforms are needed.

The main objective is to assess whether current governance frameworks adequately recognize and remedy the dignitary and identity-based harms of synthetic media, while preserving legitimate innovation and expression (Scharffs & Brown, 2025; Taeihagh, 2025).

1.4. Scope and Methodology

The study will be focussed on authentic AI-generated or AI-manipulated artificial media realistically portraying identifiable persons, particularly deepfake images or videos, fake audio or visual recordings, cloning and non-consensual intimate synthetic media. Only taken into consideration where relevant, Text-only misinformation.

It has a legal orientation, focusing on international and regional trends, with specific regard to the EU AI Act as a prime example, but also with some comparative aspects with the U.S., UK and with African perspectives on human rights (Jimoh, 2023).

In method, it is doctrinal, qualitative and comparative with a cross-section thematic approach also. It is an interdisciplinary product of legal analysis, normative human rights-based evaluation, comparative and jurisprudential regulation and literature addressing AI governance, platform governance, and synthetic-media technology. The goal is to determine if current responses are preventive, timely, accessible and effective at delivering rectification (Leibowicz, 2026; Scharffs & Brown, 2025).

1.5. Literature Review

As generative AI increasingly becomes a government issue in general, it manifests in terms of opacity, bias, misinformation, privacy concerns, power intensification, and low accountability (Taeihagh, 2025). For the synthetic media, scholars highlight the harmful aspects at every stage of their production and distribution, citing duties of producers, consumers and intermediaries (Rayun et al., 2026).

Limited value of transparency is another key theme. Users may choose to use labels and/or provenance tools but these can be ignored, misinterpreted, removed or even distrusted. They also say nothing when asked whether the represented person was contented or whether its content is degrading or illegal (Leibowicz, 2026).

This analysis is informed by the norm of human rights. Two key issues of governance of emerging technologies are dignity and remedy, and they also relate to the other themes: privacy, equality and autonomy (Gellers & Gunkel, 2022; Iturmendi Rubia, 2024). Scharffs and Brown (2025) provide an interesting contrast in the use of dignity-enhancing and dignity-degrading uses of AI.

The money donated to the scholarships is partially intended to highlight issues related to issues of identity, reputation, objectification and gendered abuse, which are specific to deepfakes. Goodyear (2025) stresses on the injury suffered by dignitaries and the necessity of greater remedies. Jimoh (2023) continues the discussion to human rights law in Africa. Saxena and Pathak (2025) emphasize the gendered aspect of synthetic sexual exploitation and focus on the main legal harm, referring to the "non-consensual synthetic media".

In general, the literature indicates that the governance is still broken and fragmented. While general AI-governance frameworks are crucial for addressing various AI-related challenges, there is a significant gap in the timeliness between these frameworks and the rights-centric approach which considers human dignity and personal identity (Leybovich, 2026; Rayun et al., 2026; Saxena and Pathak 2025; Taeihagh, 2025).

1.6. Structure of the Research

The study proceeds in five further chapters.

Chapter 2 develops the theoretical framework of dignity, identity, privacy, autonomy, and the right to one's image.

Chapter 3 explains deepfake technologies and classifies their harms.

Chapter 4 examines international human-rights law and the EU AI Act, with comparative references to other jurisdictions.

Chapter 5 evaluates whether transparency and existing remedies are sufficient, and identifies accountability and enforcement gaps.

Chapter 6 presents the conclusions and recommendations.

Chapter 2: The Theoretical and Normative Framework of Dignity and Identity

2.1. The Concept of Human Dignity in International Human Rights Law

The concept of human dignity is an underlying tenet of international human rights law. It provides justification for rights other than self-evident such as any right to life and/or to liberty, privacy, equality and/or reputation, personality, and autonomy, which depend on it, both in terms of human beings' rights and the reason of state recognition.

Freedom, justice, and peace are based on the “inherent dignity” and equal rights of all human beings as proclaimed by the Universal Declaration of Human Rights (1948, Article 1, United Nations General Assembly). Article 1 also states that everyone is 'born free and equal in dignity and rights. It is here that dignity is considered to be intrinsic and not lost due to social standing, economic worth, technological mediation and/or public visibility.

However, in the digital world dignity will defend the person against misleading reduction to data or to digital material, against objectification and humiliation, and against manipulation. This is particularly significant in contexts where the image, voice, biometric data, and identity can be captured, produced, monetised and falsified using AI and simulated media. Addressing dignity in digital regulation is essential as it serves as a link that connects autonomy, vulnerabilities, embodiment, lived experience, and institutional responsibility, as heda argued examining issues from both perspectives of digitisation and archival approaches to cultural heritage (Kadioglu Kumtepe and Riley 2024).

Dignity, therefore, is the theoretical basis for safeguarding the rights of identity and image. An image isn't a mirroring of a person; it's linked to personality, reputation, self-presentation and control over a public image.

2.1.1. Philosophical Foundations

Dignity is firmly grounded in the concept of a ‘Sitting Strong’ and philosophy has a genuine link to it in relation to Kantian ethics. Kant says that human beings have intrinsic value in that they are autonomous and moral. Thus, a person shall be considered as an end and not a mere means to someone else's.

According to Ulgen (2017) Kantian ethics provide a human-centric and non-consequentialist framework for making assessments of conduct in a technologically

mediated environment. This is crucial because being useful, productive, popular or of economic value is not a requirement for dignity.

This framework is very applicable to AI. With the continued spread of AI systems that profile, monitor, rate, recommend, assign, recognise and generate content, laws should be adapted to make humans the subjects, not the data objects, of AI systems. Disemadi and Sudirman (2025) claim that if laws would be based on AI instead of humans, that could be detrimental to the human-focused foundation of law, as AI have no genuine autonomy and/or no moral responsibility.

So, what is later added is a sociological concept of dignity based on a Kantian philosophy that adds social recognition. Fox (2025) presents a "Social Autonomy Theory" that posits that dignity acts as a shield both for autonomy and sociality. The formation of identity is related to interaction with others; use of a person's name, image, voice or reputation without authorization or to the detriment of dignity can be an attack on dignity in terms of self-determination and/or social recognition.

2.1.2. Dignity under Human Rights Instruments

Dignity is at the heart of international and regional human rights instruments. According to the UDHR dignity is the basis of all equal and inalienable rights (United Nations, 1948). The ICCPR also says that the rights are based on the "inherent dignity of the human person" (United Nations General Assembly, 1966).

Several provisions of the ICCPR are directly relevant: Article 7 restricts degrading treatment; and Article 17 protects citizens' rights to privacy and family life, honour and reputation, which Article 19 imposes restrictions upon in respect of freedom of expression.

There are also dignity systems in place in the regions. Parents can talk about dignity based on the European Convention on Human Rights, particularly Article 3 (the right to life), Article 8 (the right to a family life) and Article 10 (the right to freedom of expression). Image and image rights along with discovery of private life and autonomy are of utmost significance under Article 8. Some of the concepts of the EU Charter of Fundamental Rights are found in Article 1, which reads: "Human dignity is inviolable and shall be respected and protected" (European Union, 2012).

According to the American Convention on Human Rights (1969), Article 11 of OAS, honour and dignity are guaranteed and again under the African Charter on Human and

Peoples' Rights (1981) Article 5 of OAFU guarantees the dignity inherent in every human being.

Kadioglu Kumtepe and Riley (2024) state that dignity is indispensable in digital governance as identifying the individual agency of the citizen with structural vulnerability.

2.2. Personal Identity and the Right to One's Own Image

Personal identity is the feature(s) by which such a person is identified as an individual, including name, appearance, voice, body, gender expression, phenotypes, biometry, history, reputation, data profile, and social identity.

Rights to image is one of the important rights of one's identity. It provides an owner with rights to control their image, including its use in capture, publication, reproduction, modification and commercial use. There are two aspects of dignity and economic dimensions to this right. The dignity aspect relates to privacy, freedom, self-presentation and guarding against the humiliation or objectification of being. There is also its economic aspect, which is the value of likeness in the commercial aspect, particularly in advertisements, celebrity culture, digital performance and in the synthetic reproduction.

In European systems, image rights are to be protected by a combination of privacy, personality rights and economic rights, whereas in common law systems the protection is framed in more eclectic terms like breach of confidence, passing off, defamation, privacy, or misuse of private information, among others (Synodinou, 2014). However, there is a common theme of likeness as an expression of personality, not "information" or "property," through these differences.

Concerns have been heightened by digital technologies. Consent, personality rights and exploitation issues are raised by holograms, virtual performances and posthumous simulations as mentioned by Heugas (2021). Deepfakes and AI-generated images are also spaces of concern that present similar challenges where the person's likeness can be manipulated to impact on dignity, reputation, privacy and autonomy.

2.2.1. Identity and Informational Self-Determination

Informational self-determination is the power of (informational) self-control over the collection, use, disclosure, and interpretation of personal information. Very related to privacy and autonomy, dignity, and identity.

Chereshneva (2025) contends that the right to informational self-determination is gradually expanding to a right to digital self-determination. This is crucial as AI systems and digital platforms enable serious power imbalance between an individual and a data-processing entity. People may have no clue about the way their information is gathered, recorded, communicated or utilized, and may only give their consent superficially.

Digital self-determination must be more than simply notice-and-consent. It demands transparency, accountability, data subject rights against manipulation and surveillance, and more rights for the data subject (Chereshneva, 2025).

IDENTITY is there and it's there to inform in the digital age! Not only does it contain pictures and sound, but patterns of data as well, like location histories, biometric templates, search patterns and algorithmic classifications. Thereby protecting identity involves two aspects: representation and datafication.

2.2.2. Evolution of the Right to One's Image

The right to image has been cultivated on the basis of the law of privacy, the law on personality rights, the law on publicity rights, and the human rights jurisprudence. The subject of image protection traditionally has been defined by three things: unauthorized use, intrusion and defamation of an image. The concept of image has become increasingly linked to dignity and identity in the time elapsed since.

Barnett (1999) makes comparative studies of the United States and Spain. The image protection, as available in the United States, evolved out of concepts of publicity rights, privacy torts, and similar related laws. In Spain, it is constitutionally related to notions of "honor," "privacy" and "image." It demonstrates that the doctrine is different across different legal systems, and it proves that both the doctrine and the concept of autonomy and identity can be impacted by unauthorized use of likeness.

European law (in particular Article 8 of the ECHR) has been significant in the recognition that taking and publishing the images of a person may invade their privacy even if he or she is known to the public. This includes showing care for the image, not just in the sense of who is using it, how, and what for, but also how to keep it secret.

For Synodinou (2014) image rights is a bi-fold concept, involving the dignity as well as the economic value of an image. This may be problematic, however, in terms of copyright, media freedom, artistic expression and public information.

These are complicated further by the digital media. Heugas (2021) introduces holograms and posthumous digital performances and similar, while Prasad and Jain (2025) state that deepfakes can be an invasive manner of producing false content to threaten privacy, reputation and personal integrity. Coppo (2024) further notes that, when interacting with avatars in the metaverse, they may be extensions of their personal identity such that harm to an avatar is also harm to the dignity, autonomy or reputation of the person wearing the avatar.

2.3. The Intersection of Autonomy, Reputation, and Fundamental Rights

There is a number of fundamental rights that should be balanced regarding dignity and identity protection: autonomy, privacy, freedom of expression, artistic freedom, freedom of journalism and public interest, and rights of reputation as well as image rights.

Unauthorized use of the image of another person can infringe upon the autonomy of the person by taking away their control over the presentation of themselves. It can be disrespectful if used in an inappropriate or unfavourable manner. Could be the violation of dignity if the person is objectified, manipulated or used as an instrument. Meanwhile, the limits on the use of images could have an impact on freedom of expression, particularly in the media, art, public debate or historical archives.

International human rights law thus calls for a concept emerging from the principle of proportionality. The ICCPR provides that limitations to freedom of expression are legitimate grounds for protecting the rights and reputation of others, however the restrictions must be lawful, necessary and proportional (United Nations General Assembly, 1966). Iturmendi Rubia (2024) also highlights the points of legal, legitimate purpose, necessity, and proportionality in responding to the harms to digital dignity.

Automating harms, scaling them up, and making it hard to trace them, makes balancing more challenging in the AI context. Even though the information isn't published by anyone, one's reputation and the ability to act autonomously can be influenced by way of facial recognition, biometric profiling, synthetic media, or algorithmic recommendation systems. Prasad and Jain (2025) believe that artificial intelligence-based surveillance and deepfakes pose a risk to privacy, free expression, reputation, and personal integrity.

Dignity plays the role of a normative document, as it focuses on the balance process during this proceedings. It does not take precedence over other rights, but rather is used

to remind readers that the person has no right to be turned into a mere object of observation, entertainment, profit, being scraped for data and/or dataed, or manipulated. Kadioglu Kumtepe and Riley (2024) have noted that dignity is effective because it is tied to expanded concerns through the lens of individual lives and digital regulatory issues.

2.4. Summary of Chapter 2

In this chapter, the central theoretical and normative basis of the protection of personal identity and image rights was laid down: human dignity. International human rights law places its emphasis on dignity which is inherent, equal, and exists with or without social status, economic valuation or technological mediation. The chapter detailed the Kantian ethics' ideas; and current theories about the social autonomy, which meant that individuals would no longer be seen as objects of surveillance, manipulation, data collection or commercial exploitation. Human rights instruments, such as the Universal Declaration, ICCPR, EU Charter, ECHR, American Convention and African Charter, uphold the principle of protection of privacy, honour, reputation, autonomy and freedom from degrading treatment.

The chapter also showed that personal identity involves more than name, image, voice, body and reputation, but also data profiles, biometric information, online identities and algorithmic identity classifications. Thus, the right to one's own image entails control over the capturing, publishing, modifying, reproducing and commercial use of one's image. This right encompasses both a dignitary aspect and an economic aspect and arose in privacy law, personality rights, publicity rights, and in human rights jurisprudence.

Digital technologies and Artificial Intelligence have amplified, and exacerbated, identities related harms. The scope and intensity of harms to identity have increased with digital technologies and Artificial Intelligence. Biometric profiling, posthumous simulation, avatars or holograms, facial recognition and deepfakes have the potential to affect the dignity, privacy, reputation, autonomy and integrity of individuals. This means that self-determination is becoming self-determination over information, which needs to become digital self-determination if it is to take digital form. Thus, informational self-determination must become digital self-determination which is to be realised with recognizable transparency, accountability, consent and protection against surveillance and manipulation.

Lastly, the chapter highlighted the need for balancing identity and image rights with freedom of expression, freedom of journalistic information and artistic freedom, and public interest. As regards such conflicts, they have to be examined under the lenses of ‘legality', ‘legitimate purpose', ‘necessity' and ‘proportionality'. Nor does dignity automatically precede any other rights; it exists as the normative compass that makes it so that technological innovation and expressive freedom does not turn human being into a means of profit, being observed, entertainment or manipulation.

Chapter 3: The Weaponization of Synthetic Media: Typology and Harms

3.1. Technological Mechanics of Deepfakes and Biometric Cloning

Synthetic media can be made by AIs through artificial generation and/or manipulation of digital images, videos, audio content, avatars, and other likenesses. The most serious are probably deepfakes which can realistically portray individuals saying or doing things that they didn't actually. Multimodal AI and text-to-image have opened the doors for easier and more accessible deepfake creation for the average professional user (Hawkins et al., 2025).

New technologies like LoRA can offer parameters that enable personalized deepfake models that are trained with minimal amounts of data and consumer-level hardware. Based on these results Hawkins et al. (2025) demonstrate that model variants can be created with just 20 images, 24GB of VRAM and approximately 15 minutes required for processing. They also described an almost 35,000 publicly available model variants ready for download, and millions of downloads since the end of 2022.

All these developments are closely related to biometric cloning and digital identity replication. It's no only possible to reproduce a person's face, but also its voice, gestures, behavior and identity cues through a synthetic media. AI avatars and digital doubles are emotionally infused extensions of self—misrepresented images can result in significant identity vulnerability, as described by Poovarsan et al. (2026).

Realism, and even more automation, scale and circulation, are the principal threats! Digital environments, when considered with the lens of the amplification of harm, through anonymity and automation, enable abuse to be repeated and amplified across platforms, De Silva de Alwis (2024) contends. Therefore, wherever synthetic media is used, it is a tool of manipulation, exploitation and coercion.

3.2. Typology of Deepfake-Induced Harms

Deepfakes cause a wide variety of injuries. These encompass sexual abuse, dignity transgressions, reputational harm, identity theft and monetary exploitation, political disinformation, as well as trust damage (Williams, 2025; Li et al., 2026).

While the research into epistemic harms and the distortion of information has received a plethora of attention, much of the literature overlooks dignity-based harms, particularly consensual intimacy imagery created by AI, which represents a class that is much more prevalent, according to Li et al. (2026). This is important because even if the viewer were

an expert on recognizing that the depicted figure was a synthetic, they can still suffer serious harm from viewing such an image.

A comprehensive typology must therefore consider three levels of harm: individual; collective; and intimacy related harms – identity appropriation, financial harm, reputational harm, social silencing, harm to democratic discourse and epistemic trust.

3.2.1. Non-Consensual Intimate Imagery and Gender-Based Violence

Some of the most damaging uses of synthetic media is non-consensual intimate imagery (NCII). Synthetic NCII can be the creation of sexual content of a person even without their actual intimate picture. It differs from 'normal' image abuse as synthetic NCII can generate sexual images or videos of a person without a real intimate picture. This increases the number of possible victims and eliminates requiring the prior intimate contact (Viola & Voto, 2023) necessary for being a victim in some cases.

Empirical evidence reveals the gender-based aspects of this mistreatment. Hawkins et al. 2025 identified that 96% of publicly available Deepfake model variants made targeted towards women, and many of these were formed for NCII. However, just 29 of these apps – an analysis conducted by Williams (2025) – reveals how deepfake technologies of sexual elements, and specifically non-consensual sexual deepfakes, are designed, marketed and monetised in a way that normalises their use.

In their survey of over 16,000 individuals in 10 countries, Umbach et al. (2024) identified two reported incidents involving deepfake pornography: first, 2.2% of respondents were the victims of deepfake pornography, and 1.8% were the perpetrators of deepfake pornography. It is evident that existing legal intervention is not effective enough, since these behaviours are occurring in countries with specific laws against it.

Synthetic NCII incidentally is the latest term by which I am reading is treated in the literature, as being a gender-based violence enshrooled in technology. Digital GBF stifles women and transgender voices and restricts equal voice online, writes Dunn (2020). De Silva de Alwis (2024) defines digitized violence as a novel form of gender-based violence that is related to structural inequality. Mishra and Dash stress that non-consensual deepfake use in Asia is disproportionate towards women and girls as part of a larger issue of digital consent.

It is important to note that as long as the content is fake, the harm is not being reduced. Li et al (2026) highlight that there can still be adverse impacts on dignity, autonomy and

social status although the nature of the synthetic NCII is different and can, in fact, be more humiliating due its scale and novelty.

3.2.2. Identity Theft, Defamation, and Economic Exploitation

Uses of synthetic media are also possible for identity theft, defamatory misrepresentation and economic exploitation. Deepfakes can mimic a person's face, voice, or digital identity for false association or impersonation, or generate revenue from unauthorized use of a person's likeness.

Earlier, Poovarsan et al. (2026) have discussed the psychological implications of digital doppelgangers, suggesting that they can be perceived as extensions of oneself. Such fake faces manifested in a public environment may result in emotional displacement, loss of control and psychological intrusion.

Intimacy, criminal behavior, or deceptive social behavior can all be examples of deepfake defamation. Regardless of the false nature of a deep fake, it can cause reputational damage that is ongoing, since deep fakes may circulate on the internet and it is hard to remove them once they are there. Deepfake harms is not only deceiving viewers, but also taking and hurting the identity of those they purport to be, Li et al. (2026) stress.

It is no less also at the center of economic exploitation. According to Williams (2025), abuse is a boon to the bottom line for the rapidly expanding AI undressing apps industry. Similarly Hawkins et al. (2025) demonstrate that deep fake model repositories can be exploited on a large scale. Abuse is not only inter-personal, increasingly it is automated, platformed and monetized.

One cleavage in remedial measures, discussed by Li et al. (2024), is that reports of nudity not by consent on social media platforms created by AI technology generated no removals for more than 3 weeks as opposed to 100% removal within 25 hours in cases involving a copyright claim. This implies that platforms are more likely to react better to property claims than dignity related harm.

3.2.3. Political Disinformation and Democratic Discourse

The fake media poses threats to democratic communications too. Deepfakes can be used to generate political speech, embellish events, create fake information and undermine evidence sharing conventions. Deepfakes can be used for creating political speech, enhancing events, spreading fake information and reducing the trustworthiness of the evidence sharing conventions.

According to Ibrahim and Attia (2026), the scale of impact by AI-generated disinformation increased during the last presidential election in 2024, and social media platforms increased in their role in spreading deepfakes and misleading narratives, despite them being less used in 2016 than they had been in 2008. They also also link an AI-driven disinformation campaign to foreign meddling and cyber war strategies.

The use of deepfakes can promote polarization, misinform citizens, undermine trust in authoritative sources and in the media and in the electoral process, and compromise the credibility of evidence. This study by Das et al. (2025) demonstrates that youth who are digitally savvy are also not always able to detect fake or synthesized media, leading to a loss of trust and vulnerability to misinformation.

Yet, Don't forget that, according to Li et al. (2026), the dangers of deepfakes cannot be limited to mere political deceptive. Nevertheless, there are serious epistemic harms that need to be highlighted, but there are more frequent harms (e.g., AI-generated NCII) that also do not deserve to be forgotten.

3.3. Psychological and Social Consequences for Victims

The impact of deepfake abuse can lead to psychological and social long-term damage, such as loss of control over self and identity, reputational insecurity, fear, hypervigilance, shyness, anxiety, humiliation, and social withdrawal. Synthetic content is copied and rediscovered endlessly and so victims may suffer a continuous (not temporary) abuse.

The technology-enabled GBV brings with it safety concerns, invasion of the right to privacy, economic burden, and silencing. (Dunn, 2020). De Silva de Alwis (2024) also states that exclusion of women and girls from abusive digital environments effectively mean that they are excluded from modern public life.

Williams (2025) substantiates that AI undressing apps normalise objectification, disrupt women's privacy and body autonomy. While the subject of deepfake is discussed as a ‘virtual’ issue, deep fake messages have very real legal, social and cultural implications for women and girls, Mishra and Dash (2026) affirm.

Along with physical harm, fragmented identities cause psychological harm. Techno-emotional displacement, when digital doubles are out of one’s control, is delineated by Poovarsan et al. (2026). The victim may feel that his/her face, body, and/or personality are not his/her anymore.

The harms can be exacerbated by failing platforms. A lack of effectiveness or delayed response in reporting systems can lead the victim to remain vulnerable for a prolonged duration, as shown by Li et al. (2024). The responsibility imposed on a victim to locate, report, document and eliminate harmful content imposes further harm, and fails to recognise that the subjects have been injured. The power and pressure responsibilities placed on victims to identify, report, document and remove harmful content is a further injury and an institutional form of neglect of injury to the subjects.

3.4. The “Liar’s Dividend” and the Collapse of Epistemic Trust

Synthetic media also challenges ‘epistemic’ trust, the shared level of trust in visual and audio evidence. Deepfake technology not only creates fake content but also enables the dissemination of fake content to be dismissed as fake. This division is called the “liar's dividend.”

Trust among young users has been found to be reduced because of synthetic media as illustrated on Das et al. (2025). Knowledge of deepfakes could actually increase misinformation risk and general mistrust when users know what deepfakes are and aren't sure how to tell if they're a deepfakes or not.

Ronaldia Viola and Voto (2023) suggest that deepfakes can impact the special epistemic and emotional value accorded photographs and videos. Without the privileged link to reality of visual evidence, the notion of “seeing is believing” weakens.

This has implications for victims, institutions and democracy. Authentic evidence that fits in with abuse can be marginalized and fabricated evidence presented as plausible evidence of abuse. Certain institutions like journalism, law enforcement and courts are seeing increased uncertainty surrounding the evidence. Ibrahim and Attia (2026) illustrate how disinformation fueled by AI can capitalize on that confusion in democratic settings during elections.

Meanwhile, Li et al. (2026) caution us to not overlook dignity-based harms because of our concern for epistemic trust. Synthetic media weaponisation is not limited to the liar's dividend; it is a much more complex operation. The deepfakes undermine shared reality and impact individuals in a direct manner who have had their identity obtained.

3.5. Summary of Chapter 3

This chapter covered the processes by which synthetic media and specifically deepfakes is now becoming an ever more easily created and scalable medium of manipulation, exploitation and abuse. In recent dramatic developments, new AI systems for multimodal generation, low-cost fine-tuning, and more have enabled to produce highly realistic fake images, videos, voices and digital identities with minimal data and common hardware. In effect, synthetic media now not only allows for visual forgery, but for further cloning and replication of biometric data and identity.

Finally, the chapter demonstrated that deeper impacts than misinformation include the perils of deepfakes. Two types of harms need to be reflected —the epistemic harm and dignity-based harms— including non-consensual intimate imagery, identity theft, defamation, reputational harm, economic exploitation, and loss of trust. AI-generated intimate images, in particular, prevalent and worst form of abuse disproportionately affecting women and girls was one of the externalising technologies form of gender-based violence. This is synthetic content and it does not reduce the impact it has on dignity, autonomy, privacy and social standing.

The talk also emphasized the proliferation and monetization of forms of deepfake abuse on the platform. Synthetic media affords the entities the chance to mimic, adopt, and monetize people, with platforms' reporting systems consistently being inadequate to address the victims of harmful use. This raises the property protection concern in relation to the dignity and identity protection concern in quite large proportions.

Chapter discussed synthetic media for disinformation, and erosion of democratic discourse at the collective level. Deepfakes can create misleading information, add to confusion, destroy faith in information, news and institutions. The “liar's dividend” only exacerbates this issue, as genuine expertise or information may be called fake information, creating even greater uncertainty as to what is valid.

In general, the chapter presented the thesis that the weaponization of synthetic media takes place on a personal and societal scale. It damages people's dignity, self-esteem, and mental well-being and degrades the reliability of evidence, undermines the norms of digital and social life, and goes to great lengths to sow distrust of the system.

Chapter 4: International Regulatory Frameworks and the EU AI Act

4.1. Existing International Legal Instruments and Their Limitations

There is a fragmented, international legal framework for regulating AI, synthetic media and deepfakes. Prior to dedicated AI-specific legislation, harms arising from algorithmic systems have mostly been addressed through the different lenses of human rights law, data protection law, consumer protection law, cybercrime law and sector-specific laws. These frameworks are still relevant as AI can impact dignity, privacy, equality, free expression, access to remedies, and protection from exploitation.

These tools are, however, not built for autonomous, generative, adaptive and large scale AI systems. This makes them frequently been at a disadvantage of tackling rapid, opaque, scale, and cross-sectoral harms of AI. Deep learning copied by fake personalities does not just allow users to generate realistic images, videos, audio or text—and then use them for misinformation, defamation, sexual abuse and impersonation—but can also enable fraud, impersonation, sexual abuse and reputational harm (Fernandez, 2021; Romero Moreno, 2024).

Historical or reactive lawsuits like those claiming defamations, privacy or copyright violations, or even criminal offences are typically slow and cumbersome. They might be inadequate when damaging material moves from one jurisdiction to another and is spread around like wildfire. When content is created, hosted and distributed worldwide, victims may also be unable to identify perpetrators, prove or reason out intent, or establish jurisdiction.

Afghan (2025) identifies regulatory gaps, accountability and coordination issues, and technical opacity as key issues in international AI governance. These problems are particularly grave in the situation in which AI systems are used as “black boxes” – in which the origin of harmful output is not apparent.

Thus, while there are some important international instruments that can create a good foundation for the norms of AI governance, they do not yet constitute a complete global framework for AI governance. Firstly, their significance is in introducing principles such as dignity, accountability, transparency, non-discrimination principles, privacy and access to remedy. Such principles have been incorporated into more recent regional schemes, particularly the EU AI Act, which seeks to make these principles grounded in legal obligations for the parties.

4.1.1. Applicability of General Human Rights Treaties to Algorithmic Harms

Human rights are relevant through general human rights treaties, where AI systems can impact upon protected rights. DeepFakes can infringe privacy and reputation; synthetic sexual imagery can compromise dignity and bodily autonomy; algorithmic discrimination can impact equality and access to reliable information; and manipulative content can undermine access to reliable information.

Articles 8 and 10 of the European Convention on Human Rights are significant in Europe. Article 8 is about private and family life, Article 10 is about the freedom of expression. However, if the requirements of disclosure or content restrictions influence legitimate users, ideally in a privacy-centred approach, alongside freedom of expression, journalists, artists, creators, and AI developers, then deepfake regulation should be designed to preserve balancing exercises between known limitations on serious harm and the freedom of expression and privacy rights of others, particularly those of deepfake content creators, highlighted by Romero Moreno (2024).

Human rights law is also applicable as AI regulation can give rise to concerns on the jurisprudence of rights. Labeling, takedown systems, and any limits on synthetic content can put individuals at risk of being a victim of abuse and manipulation, while also stifling free expression. So the real question is how to design the regulation in a rights compatible manner. Romero Moreno (2024) says this is because there are more specific obligations, clearer rules and procedures that are necessary to protect rather than blanket restrictions.

General human rights treaties have their limits, however. First, they principally impact states, whereas numerous harms from AI are perpetrated by private developers, platforms, deployers, and users. Second, the scale of potential harm to human rights may be systemic, large scale and repetitive, the human rights litigation is typically ad hoc and case driven. Third, human rights principles are often vague and don't offer a set of technical guidelines concerning watermarks, labels, audits, training data, risk classification or downstream deployment.

Algorithmic harms can also occur in a cumulative and probabilistic manner. Repeatedly seeing AI-generated profiles, content, or decisions may give rise to more widespread harm, regardless of whether that harm is a clear violation of a right to privacy or freedom from discrimination. Because a single profile or piece of AI-generated content or automated decision isn't clearly infringing on a right, repeatedly seeing this content,

profiling, or automated decisions could be creating a wider harm. The responsibility of AI could be distributed among AI developers, AI providers, deployers, users, and regulators in a clearer manner, according to Afghan (2025).

This means that, despite their importance, human rights are not enough to be meaningful, and cannot delve into every aspect of every person's life. They lay the moral and constitutional foundations for the governance of AI, but must be supported by AI-specific legislation, technical standards, transparency mechanisms and institutional oversight.

4.1.2. UNESCO Recommendation on the Ethics of AI and Council of Europe Initiatives

One of the most significant soft-law documents to have been adopted at the global level is the 2021 UNESCO Recommendation on the Ethics of Artificial Intelligence. It encourages human rights and human dignity-based values with fairness, inclusion, transparency, accountability, sustainability, human oversight and a human-centered approach.

According to Ramos (2022), the aim of the UNESCO Recommendation is to orient the development of AI towards inclusion, transparency and accountability, as well as towards the rule of law, particularly in the area of work, skills and social participation. UNESCO's strength is that it has a universal scope and is endorsed by all UNESCO Member States and it is an important reference on which national AI policies will be defined.

UNESCO's idea also incorporates the linkage of AI governance to other social institutions. The use of Collective values of democracy and human rights are highlighted in innerarity (2024), a discussion on the impact of AI on Collective decision-making and public institutions. However, as Mochizuki et al. (2025) point out, UNESCO's policy guidance for the use of AI in the education sector might embody internal conflicts between notions of ethics and techno-solutionist views.

UNESCO Recommendation is the limitation as it is not binding. It does not impose enforceable duties upon states, developers or deployers. It also does not make specific provisions regarding the duties of labelers, watermarking, compliance or sanctions regarding deep fakes. The practical impact will rely on the national implementation, institutional capacities and political willingness.

In contrast, the Council of Europe's approach has become more of a legalistic one. Its efforts on AI, human rights, democracy and the rule of law aim at achieving an

enforceable framework based on fundamental rights. According to Afghan (2025), this is part of a new international accountability framework that is also complementary to the EU risk-based approach.

Council of Europe programmes and activities have difficulties, though. They are effective only when ratified, implemented and homed into States. They have to fit into mechanisms and/or processes of EU legislation, national strategies and sectoral systems. Essential treaty-level guidelines must now be put into practice at the technical and operational levels.

However, although UNESCO and Council of Europe efforts are important to provide ethical and legal principles, they cannot be a substitute for more explicit regional control, including the EU AI Act.

4.2. The European Union AI Act: A Risk-Based Approach

The EU Artificial Intelligence Act is the first EU regulation to establish a risk-based approach for AI systems and take a cross-sectoral approach. It distinguishes between different types of AI systems based on the level of risk they pose and sets obligations on the basis of the risk.

The Act distinguishes between:

- **Prohibited AI practices**;
- **High-risk AI systems**;
- **Limited-risk systems subject to transparency duties**;
- **Lower-risk AI applications**.

Cros and Thiébaut (2025) state that such an organisation has significant consequences for businesses as the requirements pertaining to compliance may impact on innovation, competitiveness, consumer confidence and even the competitiveness and viability of SMEs. The model aims at ensuring safety and respect of fundamental rights, and at maintaining technological innovation.

EU AI Act is a co-regulatory approach, where both public and private sector take part in regulation of AI (Justo-Hanani, 2026). Risks must be assessed, documented and managed by providers and deployers and compliance monitored by public authorities. This model can be useful and enrich them with expertise from the industry, but it must have good public capacities, transparency and data-sharing to avoid weak oversight or regulatory capture (Justo-Hanani, 2026).

transparence is one of the major tenets of the Act, particularly in the context of generative AIs and synthetic content. The latter, however, presents the arguments that the Act has established an uncoordinated "transparency zones", each with its own obligations, exceptions and responsibility sharing issues, hinges upon (Makauskaite-Samuole 2025). However, basic transparency obligations may be insufficient to achieve adequate safeguarding of fundamental rights, particularly as each participant of synthetic content has a distinct role in its production and dissemination by the provider, deployer and user. The AI Act, to be sure, is a significant new regulation, but it will be vital for enforcement, technical regulation, interpretation, and periodic review. However, with rapid advancements in generative AI and the nature of the threat continually changing, fixed vocabulary models might need to be updated if the obligation to the risk needs to be shifted.

4.2.1. Classification of Synthetic Content and Generative AI

The EU AI Act has regulations regarding the use of synthetic content and generative AI. Generative AI can create text, images, audio and video that can appear human-created or from the real world. The critical thing about deepfakes is that they might present persons, objects, place, entities or event realistically in a false manner.

An important difficulty surrounding the concept of a deep fake is defineditation as discussed by Fernandez (2021). Deepfakes typically employ artificial intelligence in making alterations, but their definition of legality is still under dispute. One major issue is separating the cheap, the ordinary, the parody and the satire from the legitimate in the area of visual effects and genuine artistic output and work. A benign definition can let misleading material go over the border, a lax definition can be seen to attempt to regulate legitimate creativity.

Earlier, Łabuz (2025) proposed that the phrase "existing" in the definition of a deepfake provided by the AI Act should be modified. A narrow and therefore limited definition would refer to synthetic content that depicts identifiable pre-existing persons, places, objects, entities or events. This can make it difficult to reconcile the fully synthetic but realistic content without having a reference material that parallels it, such as a PIN code, which is provided by the outdated standards existing prior to 2025. For this reason, Łabuz (2025) suggests a teleological interpretation (reading the definition through the lens of

the Act's protection); the content is fully synthetic and realistic but does not rely on any pre-existing entity or material to ground it.

Furthermore, Meding and Sorge (2025) also contend that the definition of deepfakes in the AI Act is still not entirely clear. Modern digital images can be subject to computational processing from the moment of their capture, ranging from correction and enhancement through automated image processing, and AI-assisted reconstruction. This makes it hard to tell whether and when an image might be synthetic or substantially edited.

According to Romero Moreno (2024), AI systems designed for malicious deepfakes ought to be considered high-risk systems. This would move the focus away from content, towards system design, intended use and reasonable foreseeable misuse. It would also pertain to systems creating Impersonation, Sexual Abuse Material, Fraud, or Manipulation.

Overall, synthetic media does not appear to be exempt from the duty of transparency under the EU AI Act. However, there are key questions to get a better understanding of the meaning of the Act. How these terms are defined, including “deepfake,” “substantial editing”, “existing”, and “misleading authenticity”, has a bearing on its efficacy.

4.2.2. Detailed Analysis of Transparency, Disclosure, and Labeling Requirements

A core pillar of the EU's approach to synthetic media in the AI Act is the emphasis on transparency, disclosure, and labeling. These obligations seek to give users warnings as to their engagement with AI systems and their receipt of content that has been, or is intended to be generated by, AI.

Disclosure obligations apply to deepfakes, meaning the content either must be identified as synthetic, manipulated or otherwise, or there are exceptions. Aim of reducing deception and increases visibility to users, platforms, regulators and persons affected by AI-generated content. Fernandez, though, cautions that simply printing a label isn't enough, particularly if synthetic content proliferates and/or users overlook, misunderstand, or are skeptical of labels.

One of the practical ways of transparency is watermarking. According to Rijsbosch et al. (2025), watermarking is the process of adding information to the material, denoting whether it has been created by an AI system. The empirical investigation reveals that implementation is still underdeveloped, as a minority of AI image generators employed

sufficiency in either the use of adequate watermarking or the deepfake labelling practices used. This brought to light a mismatch between the expectation by law and its technical execution.

Cousineau et al. (2025) propose a transparency card which would show the important information in an easily accessible manner. It would add more to a basic “AI-generated” label, as far as a basic notice goes, with pieces of information regarding purpose, limits, risks, data sources, human oversight and pass through accountability channels for AI generated content.

The duties of transparency involving the AI Act are however, uneven. Makauskaite-Samuole (2025) states that the transparency framework provided by the Act is not coherent as obligations differ depending on the type of AI and conditioning clauses are incorporated. Responsibility sharing is particularly challenging in generative AI because multiple actors can produce models (the so-called providers), users can create content, and the deployer can preintegrate the models into an application. It is then not necessarily defined who is responsible for the label, disclosure or watermark.

Plescak's 2025 findings extend the indefinite use of generative AI and harmful content. Further, the author suggests that the AI Act could be deficient in provisions dealing with verifying the accuracy of outputs generated by AI systems or holding individual AI models accountable in cases where they are utilized to disseminate disinformation or infringe on human rights. These responsibilities should be complemented with other EU norms (such as the Digital Services Act), and enhanced mechanisms for verification and enforcement.

Romano Moreno (2024) suggests two changes: there should be an obligation to structure the synthetic data for use in detecting and characterizing deepfakes, and AI systems designed to create deepfakes for malicious purposes should be classified as high-risk. Rijsbosch et al. (2025) agree that it is necessary to make the laws and technology governing watermarking and labelling more precise.

So, transparency is a precondition but not sufficient. While disclosure and labeling can help to limit deception, it is impossible to eradicate all harm when labels are removed, context is distorted, platforms change over time, and naïveté exists among consumers. There has to be transparency, a risk assessment, technical standards, platform responsibility, human oversight, accountability, and remedies to be effective.

4.3. Comparative Perspectives: US, UK, and Emerging Regulatory Models

There are multiple approaches to the regulation of AI and deepfakes, as indicated by comparative analysis. The EU has taken a holistic and horizontal approach towards its risk-based approach in the AI Act. The United States has taken a more sectoral and fragmented approach while the United Kingdom has preferred a principles-based and regulator-led approach.

Fernandez (2021) mentions that the EU has called for transparency, while actions taken in the United States and United Kingdom have been much more specific when addressing problematic applications of deepfakes, such as non-consensual intimate imagery and malicious impersonation. A part of it, an overall distinction, is between EU's prescriptive regulation of AI systems and transparency (ex Ante) and the regulatory approach of others, which focuses more on criminal law, tort law, platform policies and the enforcement thereof (Ex post).

AI risks can be mitigated through a sectoral approach, given that various agencies and legal regimes tackle AI related issues in sectors like consumer protection, employment, privacy, finance, health, and criminal misuse. Meanwhile, Afghan (2025) notes that sectoral regulations can lead to fragmentation and uncertainty, particularly for companies that cross border. Lacking a comprehensive AI law, requirements could be different across states, industries, and regulators.

In the UK, a more case-specific approach is taken, with existing regulation based on regulators, not a single AI act. This could minimize compliance requirements and help specialized regulators to better meet the needs of the specific situation. It can cause uncertainty, however, where there is no clear regulator or where harms to human flourishing produced by cross-sectoral applications of AI lie between institutional mandates. The UK model might be more flexible and less harmonised than the EU model.

The models for emerging regulation are also different. Afghan (2025) compares a EU risk-based approach with the US sektoral approach and a more state based approach from China. The differences make harmonisation a difficult challenge globally and compliance even more challenging for global providers of AI services, particularly in relation to transparency, auditability, data governance, human oversight and liability.

While the EU AI Act may not have immediate application to other jurisdictions, it might have an impact via the Brussels effect – the tendency for businesses to adapt their global products and services for EU compliance. Overall, the Act could have an impact on business practices by promoting trustworthy AI, compliance frameworks, and consumer confidence. Overall, the Act could influence business practices by fostering trustworthy AI, compliance, and consumer confidence. At the same time, startups and small businesses might be required to stick with the compliance which can be a pain, and hence proportional implementation is crucial.

None of the models described here can be used alone. The EU model is rather extensive, but complex; the US model is flexible, but fragmented; the UK model is adaptive, but there is uncertainty; the emerging models show different institutional priorities. The governance landscape of AI is likely to develop more complex, with a blend of legally binding requirements, technical principles, voluntary approaches, governance aspects of platforms, certification, and global collaboration.

4.4. Summary of Chapter 4

Chapter 4 looks at the impact of artificial intelligence, synthetic media and deepfakes on international and regional law, and focuses on the EU AI Act. While there are current benchmarks in international law that offer basic norms, it proposes that these are still very weak in mitigating the specific risks generated by the contemporary AI systems.

Existing statutory and regulatory frameworks, like human rights law, data protection law, consumer protection law, cybercrime law and sector specific regulation, still continue to play an important role in protecting dignified access to remedies, privacy, equality and freedom of expression. But these frameworks were not created to support autonomous, generative, adaptive and large-scale AI systems. Therefore, they cannot always react meaningfully to the speed, opacity, cross-border and the virally dispersed nature of AI-related harms. While defamation, privacy rights, copyright laws, and criminal statutes may be used against traditional forms of AI, deepfakes can be used for fraud, impersonation, sexual abuse, reputational harm, and disinformation, and traditional counteracts are reactive, limited and time-consuming.

It is important to emphasize that general human rights treaties continue to be of high significance with regards to algorithmic harms, as described in the chapter. The use of AI solutions, and of AI-generated content, can negatively impact privacy, reputation, dignity,

equality, and access to reliable information. Articles 8 and 10 of the European Convention on Human Rights (ECHR) are particularly relevant in the European context as they are concerned with reconciling the desire to avoid harm and freedom of expression. However, human rights law has a very limited scope of application; it is largely reactive, case-specific and doesn't propose specific technical rules on watermarking, labelling, auditing, risk classification or accountability throughout the AI supply chain.

The chapter also covers soft-law initiatives and treaties, especially those of UNESCO on the Ethics of Artificial Intelligence and the Council of Europe on AI's role in human rights, democracy and the Rule of Law. The 'human centered approach' laid out in the UNESCO Recommendation is grounded in dignity, fairness, inclusion, transparency, accountability, sustainability and human oversight. The main plus point of its general and ethical nature, but the main drawback is that its binding and legally compulsory. In comparison Council of Europe's work proceeds towards a more binding legal model, but necessitates ratification, implementation and harmonization within the states to be effective.

One of the major themes of the chapter centers on the EU Artificial Intelligence Act (AI Act), a groundbreaking horizontal regulatory framework for AI. The Act takes a risk-based approach, distinguishing between those AI systems activities that are prohibited, high risk systems, limited risk systems that are subject to transparency obligations, and systems that come in lower-risk applications. In this model, its purpose is that security and fundamental rights are supressed, and yet innovation is safeguarded. It also applies a co-regulatory approach where private providers take on responsibility and are required to make the risk assessment, risk documentation, and compliance of the goods they create or market, while public authorities are tasked with supervising the enforcement of this process.

The chapter emphasises transparency as an important feature of the EU AI Act, particularly for Generative AI and synthetic content. Disclosure and labelling obligations are imposed on deepfakes and AI created media to help users understand if the media and content are AI generated or manipulated. But the chapter points to the fact that transparency requirements and expectations often are not enough. Labels can be ignored, misunderstood, stripped, and distrusted; harmful content can be quickly proliferated by malicious actors through platform dynamics and/or overseas distribution.

Synthetic content and generative AI are still a murky area legally, and technically. Ambiguities and interpretative issues surround the definitions of deepfake, significant editing, "existing persons or events", and “misleading authenticity”. The narrow definition of harm could let in unregulated harmful synthetic material, while the overly broad definition could lead to over-broad regulations on legitimate uses like satire, parody, journalism, art, or visuals. Scholars consequently suggest that the EU AI Act should be construed as protective and yet respect fundamental rights.

Water-marking, labelling, disclosure are other important mechanisms which are not consistently adopted in practice explained in the chapter. Experimental studies indicate that current AI image generators offer suboptimal Watermarking and/or labeling of Deepsays. Transparency tools, like transparency cards, can help increase accountability by providing easily accessible users with the information they need to know about an AI system, including its purpose, risk and benefits, data sources, limitations, human oversight, and complaint procedure. Nevertheless, transparency cannot be the only way to avoid all harms; along with transparency it requires the implementation of technical standards, risk evaluations, a platform's responsibility, platform enforcement, and remedy.

In comparison, the chapter examines the EU model together with regulatory strategies in the United States, the United Kingdom and new regulatory areas. EU has taken a holistic approach to risk-based ex ante framework. In contrast, the U.S. does not have an integrated approach and presumes upon non-uniformity of agencies, state laws, criminal law, tort law, privacy law, and consumer protection. This offers some flexibility, yet can be a bit unclear and inconsistent. In the UK, a principles-based approach with a focus on regulation is favored, as it is more adaptable but may fail to harmonize where two regulatory regimes contradict one another with regards to harms from cross-sector AI use.

At the end of the chapter, the authors reach the conclusion that no one model is complete. The EU AI Act is comprehensive, but also complex; the US approach is flexible, but for now poorly connected between themselves and with bodies at the EU level; the UK model is flexible, but also incredible, and other developing models show various institutional and political preferences. Enforcing AI governance will likely need a mixture of binding legislations, technical standards, voluntary frameworks, certifications, AI platform governance, enforcement, and international cooperation. Overall, the EU AI Act is a

significant step towards operationalizing key principles such as transparency, accountability, safety, and protection of fundamental rights, yet the extent to which it can truly work in practice, requires interpretation, enforcement, technological innovation, and integration with other regulatory frameworks.

Chapter 5: Evaluating the Sufficiency of Transparency and Legal Remedies

5.1. The Limits of Transparency: The “Unringing the Bell” Dilemma

One of the primary regulations to synthetic media is transparency. Some models of governance take the view that if AI-produced or manipulated content is marked, disclosed, or watermed, primary damage will not be as severe because one would be able to distinguish the content was made by a computer.

But there are some restrictions on this assumption. Very few produce transparency – it may come at the end. But once a harmful deepfake has been seen, copied, downloaded, commended, or graced with public memory, it's hard to rig it back up. However, when the victim's dignity, privacy, reputation or autonomy has been violated by a harmful deep fake, it's difficult to undo those effects once it has been seen, copied, downloaded, commended over, or incorporated into public memory. Memories of content circulations will remain firmly in our society and our mind and post-loud claims that the content is false or synthetic still cannot compensate for the social and psychological side effects.

Past disclosure frameworks, says Rockwell and Conklin (2026), have not succeeded when basic regulatory requirements have been compromised by synthetic actors, such as any actor that might not be identifiable, might not have the claimed incentive, and is incapable of being held to account for the disclosure. A label can show it's AI generated, without attributing to the frostpeter, media to prevent re-distribution, or make any monetary compensation to the victim. Perriello (2026) also holds that duties on transparency in the EU AI Act are helpful, but can fall short in the case of 'intimate abuse', 'manipulation of identity' and 'disseminating disinformation across borders', when synthetic media is involved.

Effective transparency applications regarding deepfakes and non-consensual pornography are particularly relevant. Deepfakes and synthetic non-consensual sexual imagery are noteworthy applications of limits of transparency. Such content is not only misleading, but is also attacking sexual autonomy, dignity and social identity. Spanning psychological, social, and economic effects, deepfake pornography harms victims, and the legal system offers inadequate protection and erasure in a timely manner, as emphasized by Nasution et al. (2025). Saenz claims that tort and IP laws offer hindersight remedies in the case of deepfakes, while non-consensual pornography laws offer no remedy where women are the victims of sexualized identity abuse.

Hence, the word “synthetic” or “AI-generated” isn't the answer to feelings of humiliation, fear, or reputational damage, or loss of control. Being transparent may deter deceit but it certainly doesn't heal dignity-related injury.

5.1.1. Does Disclosure or Labeling Restore Violated Dignity and Reputation?

Disclosure and labeling are tools that can help audiences to know the difference between authentic and synthetic content, but they cannot or should not restore dignity or reputation. However, deepfakes aren't just a deceptive threat. Also it is an intrusion upon one's identity, violation of privacy, emotional distress, reputational harm, and autonomy is being withdrawn.

Nowak et al. (2023) explain that deepfakes pose a threat to legal systems for the following reasons: deepfakes manipulate identity, defy existing defamation, privacy, tort and criminal law doctrine and can fabricate audiovisual evidence. Subsequent label can correct the factual record, but will not take away from the original reputational harm.

It's a distinction that is crucial. Disclosure may be sufficient to eliminate epistemic deepfake harm. However, there are numerous dignity harms. The individual is at a disadvantage because of the appropriation without consent of his or her face, voice, body or identity. The impact on dignity, autonomy, moral agency, personality rights and control over a person's digital identity has been highlighted by Purcell et al. (2026) in the context of personal AI replicas.

The same view is taken by Michałkiewicz-Kądziela and Staszewski (2026) with regard to the right to be forgotten. They say that GPDRH Article 17 doesn't work in today's deepfake age due to the ability of malicious synthetic media to be copied, shared and repurposed through various platforms. They say that their goals are to shift from the right to “erasure” to the right to control an individual's digital identity.

Unfortunately, when the harm is not in sight but rather assumed, labeling also proves ineffective. Mahalaldaya notes that in the case of deep fake pornography, even awareness that the content is synthetic, does not prevent users from experiencing sexual humiliation, objectification, harassment, archival, and/or redistribution. In the age of AI, Patil and Mishra (2026) contend that "reactive takedown models" are not sufficient, as deepfake pornography falls short of dignity, privacy, and sexual autonomy.

So, transparency is no replacement for removal, injunction and compensation, and holding platforms accountable.

5.1.2. Cognitive Bias and the Viral Spread of Labeled Synthetic Media

The problem is compounded by the fact of cognitive bias and virality of the platform. In the event, audiences will likely still recall the erroneous belief if it was originally conveyed synthetically. In fact, when the information is emotional, persuading, or often compatible with previous beliefs, these corrections are not always successful at eradicating previous beliefs.

Audiovisual content is also made to seem real, making this a more acute issue with the Deepfakes. The errors in the outputs of AI are epistemic challenges, as, according to Poibeau (in press), users are prone to believing that machines' said information is reliable, grounded, or authoritative. One might think that artificial art will never be convincing, but in synthetic media, that is not true and visual realism can continue in the face of obvious artificiality.

This is known as “synthetic authority”, says Rockwell and Conklin (2026), as credibility is created through algorithms, artificial identities or “personalas.” When a viral deepfake first shows up, it can be shared, commented on, reposted, even amplified by influencers and may appear to be credible even with the warning label attached.

There is also a lack of transparency on platforms, as media that is fake can outpace the efforts to verify, moderate or even respond to it with legal action. When looking into the Almendralejo scandal Somotomà (2026) reveals a harm pathway: model development, model output at application level and model dissemination at platform level. Notice and takedown is ineffective when content can be redistributed prior to its being moderated.

Content, regardless of those with a label, can be downloaded, screen recorded, cropped, translated, reposted or transferred to encrypted channels. So it may not be useful to label after circulations.

While watermarking and authentication tools can aid in solving the problem, they do not entirely solve the issue. Many of the image watermarking systems based on AI can be easily removed and are not interoperable between various platforms as found by Deshpande and Kanti (2026). They suggest more robust systems based on digital rights management, machine-learning classifiers, and distributed digital signatures recognised by blockchain. Even secure watermarking can, however, not completely correct the dignity when calamity has reached publication, though.

Therefore, technical clarity instruments will be useful for identification, evidence and enforcement but only in the context of a larger legal remediation structures.

5.2. Loopholes and Enforcement Gaps in Current Governance Models

Governance is currently poor and disjointed, it is reactive and selectively applied. Deepfakes are governed by a myriad of laws, including privacy, defamation, IP, platform, criminal, data protection and AI-specific laws, to name only a handful. This division generates ‘gaps' pun intended.

Traditional legal remedies fail to deal with defamation, identity theft, and privacy breaches stemming from traditional deepfakes, says Darma et al. (2025). They seek clarification, enforcement and a complete revisiting of legislation.

The biggest problem is that laws today are only written to address specific harms and not synthetic manipulation in and of itself. Not all impressively disrespectful, inappropriate sexual or embarrassing deepfakes need fall within defamation doctrine. The principles of privacy law may apply if the information is misused, but only if it relates to a private recording – they do not necessarily apply to created "content" that is not a private recording. Copyright owners and/or copyright performers can be protected by intellectual property law, but not the general public against synthetic identity abuse. Thus, Saenz (2024) contends, traditional treatments are still incongruent and inadequate.

The other hurdle is the use of the “notice and take down” model. Such systems presume that illegal content can be flagged and taken down in time that incurs no permanent damage. However, deepfakes can be easily copy and pasted and the victim can only come to find out about them after going viral. Patil and Mishra (2026) theorize that the framework of intermediary liability in India instantiate a pre-AI model due to intermediary neutrality and takedown responsibilities.

The EU framework is more developed since it encompasses the AI Act, the Digital Services Act, the GDPR, and emerging regulations on (gender based) violence against women. But even this multi-layered system is not foolproof, Perriello (2026) claims, since the damages inflicted by deepfakes are also carried through four processes—creation, distribution, amplification, and injury. Victims may continue to face delayed, harmonious relief and assistance.

5.2.1. Jurisdictional Challenges and Cross-Border Perpetration

As with most harms, the damage wrought by deepfakes is inherently cross-border. A model can be trained in one country, then utilized in a second country, then hosted on a certain platform of a third country, and then targeted at a victim in a fourth country. This leaves an important ambiguities for jurisdiction.

The example cited by Pathak (2026) is that of the AI system that had been trained in the USA but used by an Indian user that was hosted in a Chinese platform and came to defame a French citizen. what law applies may not have any real resolution. The private international law rules are divergent and are also ‘epistemically incompatible', otherwise allowing regulatory arbitrage and victimless remedies, Pathak argues.

Quick (2026) manages to elucidate the concept of territorial jurisdiction in the context of cybercrime, as cybercrime can specifically target a place without physically going there, and can affect harm to a place without a presence there. Traditional concepts like territoriality, nationality and protective principle could be both complicated and complicated when behaviour, facilities, information, and damages are shared between states. This reasoning holds true when applying it to deepfakes too.

Civil and criminal remedies are both hampered by jurisdictional issues. A court order may be available from a domestic court; the host platform may be located outside the victim's country; the person committing the violation may be unknown; and the jurisdiction of the host platform may be one in which the mechanisms for cooperation are underdeveloped. Where there is co-operation, there can still be delays which mean remedies are ineffective.

According to Nasution et al. (2025), deep fake pornography victims in Indonesia are encountered with shortcomings in rights to be forgotten claims because of the lacked legal framework, unclear implementation and weak enforcement. The situation of victims having to seek remedies in multiple jurisdictions occurs worldwide.

Instead of following the traditional way of looking at the liabilities based on territoriality, Pathak (2026) suggests that focus should be moved to “demonstrable control”. In this approach, actors should be held accountable if they have any practice-to-practice control over the creation, hosting, amplification, monetisation, removal and mitigation of harmful AI-generated content. This is a better representation of the architecture of digital-harm.

5.2.2. The Accountability Gap Between Platform Intermediaries and Creators

Another huge enforcement gap is the lack of responsibility for creators, users, platforms, AI creators and infrastructure providers. Typically the traditional model is set up around the immediate compromiser or uploader. In addition to the deepfake technology, many other players, such as open-source models, nudification applications, hosting platforms, messaging apps, recommender tools and search engines, payment systems, etc. facilitate deepfake harm.

Liability is complex and involves creators' drop, platforms, AI developers, and malicious users, Nowak et al. (2023) stress.

A creator centered model is flawed because individual perpetrators can be anonymous, unjudgeable, a minor, out of this country, or hard to identify. In the case of the Almendalejo scandal like that of Somtoma (2026), harm was constructed through a progression of activities involving the creation of AI garments, nudification applications, and distribution over the Telegram message platform.In the scandal of Almendalejo, harm was fabricated through a series of actions from making of the AI garments, the nudification applications, and distribution over the Telegram message platform, (Santomà, 2026). The EU governance instruments in place are frequently in their failure to provide a sufficient ex ante duty to prevent foreseeable harm to those who are best placed to do so.

There are also issues of platform immunity and safe-harbour which make it harder to hold anyone accountable. It becomes a legal fiction, when algorithmic systems promote harmful synthetic content, Patil and Mishra (2026) say. Platforms are not just passive if they create designs with deepfakes prominently visible.

Similarly, Gupta (2023) suggests that Indian regulation is not enough to deal with the complexity of deepfakes and proposes for legislation that is sui generis or “special rules for deepfakes.” Such regulations must define the liability of creators as well as platforms, AI application providers and intermediaries.

If it isn't explicitly defined, then there is a vicious cycle of accountability: platforms say that they are neutral, and developers say that they only supply general purpose tools or no tools at all, while perpetrators stay anonymous and cannot be reached.

Rockwell and Conklin (2026) suggest a regulatory model that is synthetic-aware, ensuring the three pillars of identity verification, platform liability and subsequent

international coordination. This is not just financially relevant as deepfake governance cannot be confined to the task of labeling content individually. It will demand from platforms and AI service providers that they certify high-risk uses, keep record of the uses, gather document evidence and assist enforcement.

5.3. Toward a Rights-Based Legal Remedy: Injunctive Relief, Removal, and Reparations

Transparency and the existing enforcement system need to be strengthened and adopted to a rights-based remedial model for deepfakes. According to this model, instead of meaning misinformation, harmful synthetic media amounts to an assault on dignity, privacy, autonomy, reputation, sexual integrity and control of digital identity.

These remedies are supposed to prove preventive, corrective and reparative in nature.

Victims' first priority should be the equitable and prompt issuance of injunctive relief. Courts and regulators should have the authority to demand that it be removed, de-indexed, disabling access - and preservation of evidence and prohibition of further distribution. Speed is crucial as the harm of the deepfakes grows as they spread. This is because, says Santomà (2026), reactive notice-and-takedown is unsuccessful if the architecture presents a platform can enable the redistribution of content prior to any moderation action.

Second, the right of removal must be more than merely a right to takedown. Michałkiewicz-Kądziela and Staszewski (2026) suggest expanding the right to erasure to a new right to control digital identity. Michałkiewicz-Kądziela and Staszewski (2026) call for the right to erasure to be extended to a new right of controlling one's digital identity. Important as the deep fake is more apt to be a synthetic reconstruction of appearance, voice or personality than merely a copy of personal information. Remedies thus include coverage for the unauthorized representations subsequently created by the use of AI tools when there is no original intimate recording.

Third, there should be duties of care that are recognised as enforceable on platforms. These should encompass risk assessment, "fast-track" channels for response on the victim, hash-matching/fingerprinting of known harmful deepfakes, repeat upload prevention, transparency reporting, and co-operation with law enforcement agencies, where appropriate. Patil and Mishra (2026) call for a "dignity-based" approach to safe-harbour, rather than one that is passive.Patil and Mishra (2026) propose a dignity-based

approach to safe-harbour, moving beyond a passive framework to acknowledge the structural obligation of platforms to prevent AI-driven harm.

Fourth, reparations shall be available to victims as their fourth right. The loss of reputation, emotional harm, economic/savings damages, legal expenses, counseling, cyber security support, and reputation management alone cannot be replaced by the removal. According to Saenz (2024), there should be bold, clear and effective legal safeguards for victims of deepfake pornography, with criminal laws specifically addressing the harm. Consider recovery costs for remediation, compensatory damages, reasonable—statutory damages or punitive damages in serious cases.

Fifth, there must be technological authentication systems to support remedies. Deshpande and Kanti (2026) suggest improved watermarking and verification using Digital Rights Management (DRM), Adaptive Classifiers, and Blockchain verified Digital Signatures. These are the tools that can assist in attribution as well as detection, evidence preservation and enforcement. They have to be compliant with the law, and they have to be interoperable with other platforms.

Lastly, the framework for cross-border cooperation needs to be integrated in remedial design. Pathak (2026) suggests a multi-level value chain accountability framework that is driven by proof of control not territoriality. A similar approach to jurisdictional analysis, by way of functional connections, in lieu of strict territorial arguments, is espoused by Quick (2026). These methods are important for a variety of reasons, in part because most deepfake actors reside on different jurisdictions.

A victim centered approach should enable jurisdictions to assert jurisdiction when there is significant harm suffered by the victim, where a platform serves or targets victims in jurisdiction, and where a jurisdiction has some meaningful control over the harmful content.

5.4. Summary of Chapter 5

In this chapter, the authors present the findings and conclusions regarding whether disclosure requirements and existing legal protection are sufficient to neutralise harms resulting from a deepfake and other synthetic forms of media. It proved that labelling, disclosure and watermarking are effective in assisting users in determining if content is AI generated but not the complete solution to prevent or remedy the resulting harm. After the bad information is passed around, it's nearly impossible to fix the psychological,

reputational, economic and social effects. “Unringing the bell” is the issue: transparency might clear the misrecord of facts but it can't repair the damage to the victim's dignity, privacy, autonomy, or the victim's control over who he is.

However, the lack of transparency is especially noticeable in instances dealing with sex crimes related to deepfakes and sexually explicit synthetic images. Such is not always the case; in these situations, it is not the validators' credibility that causes people to suffer harm. The appropriation without consent of a person's face, voice, body or identity can be an injury to dignity, confidentiality, and sexual self-determination itself. Further, false impressions can stick with content even after it is corrected or labeled, as the "cognitive bias" of the media, "audiovisual realism," and the "virality" of the platforms. Synthetic media may also be downloaded, modified and spread on more platforms, making practical implementation of labels, watermarks and notice-and-takedown mechanisms lachese. They can also be downloaded, manipulated and redistributed across platforms, limiting the utility of labels, watermarks, and notice-and-takedown systems.

The chapter also found that an important challenge was the lack of governance structures and mechanisms. The existing laws in this area are still scattered across privacy, defamation, intellectual property, criminal, data protection, platform and AI laws. Most of these regimes tackle specific results of deepfake abuse, not the issue itself of a synthetic identity being manipulated. The weakness in enforcement also resides in reactive takedown protocols, anonymity of those responsible, ability of others to speedily pass along, and lack of definition of the duties of platforms and AI service providers.

This is compounded by crossing the border. The production, hosting, distribution, and victimization of a deep fake may take place in several jurisdictions, complicating the application of law and jurisdiction, collection of evidence, and enforcement. When it is a domestic court-ordered against an out-of-country perpetrator, platform, or data infrastructure, it may have limited impact. Thus, only territorial connection is not to be a basis of liability. Better would be a more specific responsibility placed on actors for "demonstrable control" that would be able to meaningfully avert or destroy harmful fake content, or limit, monetise or minimise its reach.

The chapter also noted a lack of accountability hackers have to the broader technology ecosystem. Potential victims of deepface harm are model makers, application developers, application hosts, social media platforms, recommender systems, search engines,

payment processors, and more. Attaching a blinders action to the original creator's actions is not sufficient since the culper(s) can be nameless, socially unsatisfied or even unavailable in the country. Platforms should not be used simply as passive entities where technology actively suggests, promotes and encourages the proliferation of harmful content.

Finally, the chapter proposed a victim-centred and rights based remedial process. This should include speedy injunctive measures, de-indexing and removal, evidence preservation, prevention of further uploads, and general safeguarding of digital identity. Enforceable duties of care should be imposed on platforms and AI providers such as carrying out risk assessments, providing easy avenues for victims to report content, adding content fingerprinting, providing traceability and facilitating cooperation with enforcement bodies. There should also be meaningful reparations, such as compensationary and statutory damages, cost of counseling, legal fees, cyber security, and reputation management fees for victims.

Technical means like watermarking, authentication systems, and digital signatures will complement and help in detection, attribution, and enforcement, but legal action will not be eradicated. An integrated governance system that ensures transparency, prevention, a platform's accountability, fast removal, compensation and cooperation between jurisdictions, therefore, is essential. The main argument is that transparency is not enough: the regulation of deepfakes should go beyond simply identifying synthetic content to preventing harm, restoring rights, and offering effective remedies in cases of deepfakes that infringe on dignity, privacy, autonomy, reputation, and digital identity.

Chapter 6: Conclusion and Recommendations

6.1. Summary of Key Findings

It has considered whether international human rights and AI-governance principles respond well enough to the harms that stem from deepfakes and other forms of non-consensual synthetic media, affording a sufficient protection of human dignity and personal identity. The key takeaway is that existing mechanisms offer important normative standards but are not yet robust enough, cohesive, enforceable, and victim-centred to address the special harms of synthetic identity manipulation.

First, the study reaffirmed that human humanity is the fundamental norm on which the protection of identity in the time of artificial intelligence is based. The dignity of each person is connected with his or her being treated as an autonomous subject and not as an object to be manipulated, surveyed, exploited, or sold by technological means. Here the concept of identity has nothing to do with legal name or physical appearance. It involves image, voice, body, biometric, digital persona and other characteristics by which a person is socially recognized and represented. Deepfakes can shape, distort, or steal these aspects of identity and directly interfere with them by creating, modifying, and spreading synthetic images of oneself.

Second, the study illustrated that there are non-informational harms within deepfakes. While synthetic media may of course deceive viewers and cause a loss in trust of evidence, this effect is also personal and from the level of dignity. Non-consensual sexual deepfakes, biometric cloning, impersonation, and identity fabrication, when used to the detriment of a person's privacy, autonomy, reputation, and/or psychological integrity, constitute breaches of those rights even if the viewer is aware that the content is fake. Deception is only part of the problem when it comes to injuries—namely, when someone's likeness, voice or identity is used for manipulation, fraud, humiliation and exploitation without consent. It is particularly important that many legal systems persist in restricting their attention to either falsehood or damage to reputation or to economic loss, rather than taking into account the moral wounding that results from unapproved “unauthorized” synthetic embodiment.

Third, the research revealed that there is still a lack of cohesive and comprehensive legislation. Synthetic-media abuse covers various areas of the law, with human rights relating to some facets, privacy law to others, data protection to certain aspects, defamation and intellectual property to others, and platform regulation to others. Reliance

on traditional legal responses tends to be too slow, more geographically restricted, and more cumbersome in nature to adequately deal with viral, geographically diverse, anonymous harms. In particular, deepfakes' harms are frequently deemed to be instances of pre-existing types of harms, as opposed to a newer type of digital identity violation that deserves specific protection.

Fourth, the study determines that the obligations on transparency are helpful but not enough. Labeling, disclosure and watermarking can increase awareness and increase the distinction between true and false content. However, their efforts fall short of preventing the creation or dissemination of harmful deepfakes and are not comprehensive in addressing the damages caused by such created content once it is spread. The ‘unringing the bell' question remains critical: although false synthetic content may be subsequently identified, its social, psychological and relational impacts can remain. Furthermore, labels could be ignored, removed, misinterpreted and become ineffective when they are virally and re-redistributed by platforms.

The fifth finding was that there is a severe accountability shortfall in the synthetic-media "ecosystem. Harm is very seldom solely carried out by the single creator. It typically requires a broader ecosystem of actors such as application providers, platforms, search engines, payment intermediaries, hosting services, and model developers and recommender systems. A one-dimensional target fixation on the original perpetrators is therefore insufficient, especially if the perpetrators are anonymous and/or unjudgeable or overseas. An effective regime will specify the actors who can be shown to have control over the production, amplification, monetization or persistence of harmful synthetic content.

Last but not least, the study has reached a conclusion that the future governance require special victim-centred, rights-based approach. The best answer to deepfake harms is a comprehensive strategy of prevention, transparency, platform accountability, early mitigation, trackability, cross-border collaboration and reparations. There are elements of such a system embedded in existing systems; particularly those related to dignity, privacy, equality and accountability. But those values need to be manufactured now into more precise, actionable commitments – commitments in the algorithmic age.

6.2. Policy and Legal Recommendations for International Human Rights Frameworks

Based on these results, the following policy and legal suggestions are offered regarding the international human rights system and some related AI-governance frameworks to improve the system's functioning. To enhance the functioning of international human rights the following policy and legal recommendations are made based on these findings: International and regional laws should explicitly address non-consensual synthetic identity manipulation, as a human rights matter. Deepfake abuse should not be considered simply a by-product of innovation, or simply a subcategory of misinformation and copyright infringement and reputational harm. International bodies, courts and policy makers should recognise that an unauthorized synthetic reproduction or manipulation of a person's likeness, voice or biometric identity may be a violation of dignity, privacy and autonomy, and human identity rights in itself. This conceptual understanding is crucial for the critical goal of appropriate law-based response to address the real nature of the harm.

Second, the right to identity should be interpreted in a technologically updated manner. The current principles of privacy, respect for the person, reputation and dignity should be extended to include digital identity, the representation of the person's biological data, the AI's impersonation of the person's identity. We need an understanding of the image, voice, face, gesture and other personal identifiers as part of the personhood that is not just physical or expressive but datafied and reproducible in digital space. The legal doctrine of information self-determination has to be broadened to a principle of digital self-determination.

Third, third parties, in particular the states, should establish clear civil and, in cases where it is applicable, criminal prohibitions against non-consensual synthetic sexual images, malicious impersonation and harmful biometric cloning. Laws of this nature need to be developed with care to ensure that they regulate abusive behaviour, but not legitimate uses like satire, artistic or public-interest reporting, or research. The legal criterion shouldn't simply be based on realness or deceptiveness, but also on non-consensuality, harm, exploitation or infringement of rights. Focus on harm and use context sensitivity is much more likely to be effective than simply being focused toward technical definitions.

Fourth, international human rights mechanisms must give effect to the duty of the platform and AI service providers. All stakeholders producing, showing, recommending and financially benefiting from synthetic media should have commensurate responsibilities and powers to avert harm. They might encompass risk assessments, measures for safe design, provenance tools, content fingerprinting, reporting systems that are accessible, repeat upload prevention systems, evidence preservation systems, and collaboration with competent authorities. A platform or provider that has a meaningful role in preventing that such harms could be meant to ensure it would not create the impression of acting merely as a passive intermediary.

Fifth, there is a need to retain the transparency obligations, however these should be included in a larger remedial framework. Labeling and watermarking systems work well for raising user awareness and for providing evidence in case of an incident, but they do not constitute, nor should they be used as a replacement for, prevention and remedy. There needs to be clearer international guidance that transparency is just one aspect of governance. It's necessary to complement the provisions on traceability, and authenticated provenance; conditions for auditability; and legal penalties for intended circumvention of protective measures.

Sixth, domestic law shall give swift and a ready medicine to the victim. Remedies shall contain: expedited injunctive relief; notice-and-removal procedures and de-indexing; removal of the repeat upload; preservation of the digital evidence; and, if warranted, an emergency protective order. There is no justification for victims to need to pursue a number of disjointed legal processes before receiving a simple form of protection. Procedural ease, complaint mechanisms, legal counsel and trauma-informed enforcement mechanisms are also key to effective access to justice.

Seventh, those who have been harmed should be given all the reparations they deserve. The compensation can not be restricted to the issue of monetary damages! Emotional harm, reputation damage, counseling fees, cyber security resources, legal fees and the ability to re-establish one's online profile and personal security should also be considered with regards to humanitarian and rights-based remedies. Statutory damages or presumptive remedies can help correct under-enforcement, especially in cases where harm to the individual is hard to measure, and decrease barriers to relief.

The eighth one is to enhance cooperation between the countries. In most cases, deepfake harms engage more than one jurisdiction; the victim's jurisdiction, the perpetrator's jurisdiction, the jurisdiction where the platform was operating, where the servers were located, and the jurisdiction where the training and/or deployment took place. States should therefore consider further improving the platforms, information sharing, obligations and technical standards on urgent orders and the capacity to provide mutual assistance. Minimal procedural protections and model legal principles, and harmonized definitions of the principles of case law, can enhance the contribution of international organisations.

Ninth, policymakers should inculcate development of trustworthy technical standards while keeping in mind the need of policy solutions other than technical ones. Traceability and detection systems can enhance with provenances metadata, digital signatures, tools and watermark systems, but they still can be removed, evaded, and unevenly adopted. These are most important in facilitating efforts to hold law enforcement accountable, in moderating platforms, and in securing evidentiary reliability. For this reason, technical standards should be set to respect human rights values and should be placed under independent monitoring.

Tenth, AI governance needs to be grounded in the principles of human rights and should be designed to meet these goals. Risks to dignity, identity, privacy, equality and autonomy should be considered at the very conception, training and release of a Generative AI system, and when implementing it, until all impacts have been mitigated. That involves examining the potential for modelling that can be used to impersonate a person, sexual abuse, fraud or misuse of the persons biometric data, as well as taking measures to prevent such misuse. International frameworks should go beyond lip service to ethics and call for action steps that can be measured, audited and implemented.

6.3. Concluding Remarks on the Future of Human Dignity in the Algorithmic Age

Deepfakes and synthetic media, precisely where they are sourced, throws many legal frameworks into a disorienting state of unease as it skewers the nexus and dynamic of human identification, representation and control. In previous legal settings, damages to image and identity typically have been associated with photography, publication, defamation or surveillance. AI, however, in the algorithmic era, can create a completely

different image of a person, whether it's their speech, actions, appearance, or their presence at events, that never had existed. This changes the nature of identity from being ‘observed' or recorded, to ‘replicated', ‘manipulated' and ‘circulated' on an artificially large scale.

Whether legal or institutional structures will be able to accommodate this shift without turning persons into data points, training material and commercially-exploitable avatars will be the determining factor for the future of human dignity. The real issue is not one of whether societies should allow for technological innovation, but of the possibility of innovation being held in respect for the human person. Dignity cannot be had seriously if the law fails to give people any meaningful control over how they are reproduced, manipulated and deployed through AI systems.

As on the other hand, the future legal order should not disregard the absence of simple solutions. There is a need for a careful balance between deepfake regulation and freedom of expression, artistic creativity, journalism, and legitimate research. Not all synthetic depictions are vicious, nor is all digital manipulation abuse. Balancing the need to specifically outlaw exploitation and coercion with the need to allow for lawful and socially beneficial uses of synthetic media is therefore a challenge in constructing legal frameworks that would achieve these objectives. There must be proportionality, examining situations in their context, and ongoing conversations between human rights law and technology regulation, and between platform governance and the law.

Finally, this study suggests that human dignity should always be the guiding principle for regulating AI. How to do good to one another in digital environments is less about how to handle false information, it is about how to handle personhood itself, that is the crux of deepfakes. Humiliation, sexual abuse, deception, and making use of a person's face, voice, or body without his or her consent—the extent of the harm extends beyond misinformation to the conditions of individual selfhood and social recognition. Hence, the dignity and identity should be at the heart of future international laws aimed at addressing the present and future effects of AI.

The algorithmic age will be seen if the human rights law can keep up with the pace with which it is evolving to safeguard what is inherently human. This success will rely on it being more than a series of disconnected and isolated responses or "canned" reactions to crisis preparedness; more than a formal process; more than being "reactive" – but more

than a mere "preventive"; more than being "national" or "international" – more than being "lived. While transparency is significant, it's not sufficient. But the future requires a more imaginative response: the preservation of human dignity must be central and not secondary concerns of technological governance, in a world of synthetic realities.

References

Afghan, V. A. (2025). Legal mechanisms for audit and accountability in artificial intelligence systems in the context of international law. *Scientific Bulletin of Uzhhorod National University. Series: Law, 5*(90). https://doi.org/10.24144/2307-3322.2025.90.5.8

Barnett, S. R. (1999). The right to one's own image: Publicity and privacy rights in the United States and Spain. *The American Journal of Comparative Law, 47*(4), 555–581. https://doi.org/10.2307/841069

Chereshneva, I. A. (2025). The transformation of the right to informational self-determination in the era of artificial intelligence and other digital technologies: Challenges and prospects. *Legal Studies, 12*, 72–82. https://doi.org/10.25136/2409-7136.2025.12.77307

Coppo, L. (2024). Fundamental rights and the metaverse: Avatar–player relationships. In L. A. DiMatteo & M. Cannarsa (Eds.), *Research handbook on the metaverse and law* (pp. 79–97). Edward Elgar Publishing. https://doi.org/10.4337/9781035324866.00013

Cors, M. S., & Thiébaut, R. (2025). Artificial intelligence and the impact of the EU AI Act in business organizations. *AI Magazine, 46*, e70039. https://doi.org/10.1002/aaai.70039

Cousineau, C., Herger, N., & Dara, R. (2025). Transparency requirements across AI legislative acts, frameworks and organizations: Shaping a sample transparency card. *AI and Ethics, 5*, 4255–4267. https://doi.org/10.1007/s43681-025-00725-5

Darma, M., Gisymar, N., Purwadi, A., Zubaedah, P., & Judijanto, L. (2025). Legal implications of deepfake technology misuse in digital content on social media. *Science of Law, 2025*, 98–103. https://doi.org/10.55284/eqazc148

Das, S. S., Agarwal, M., Rajan, S., & Padhi, A. (2025). Synthetic realities: Youth media literacy and trust in the age of digital deception. *Information Development, 0*(0).

de Silva de Alwis, R. (2024). A rapidly shifting landscape: Why digitized violence is the newest category of gender-based violence. *La Revue des Juristes de Sciences Po, 25*, 62. https://ssrn.com/abstract=4648409

Deshpande, A., & Kanti, V. B. (2026). Insecure AI image watermarking—Is it really damaging the future? In *Proceedings of the 2025 9th International Conference on Computer Science and Artificial Intelligence (CSAI '25)* (pp. 419–431). Association for Computing Machinery. https://doi.org/10.1145/3788149.3788154

Disemadi, H. S., & Sudirman, L. (2025). Reassessing legal recognition of AI: Human dignity and the challenge of AI as a legal subject in Indonesia. *Masalah-Masalah Hukum, 54*(1), 1–12. https://doi.org/10.14710/mmh.54.1.2025.1-12

Dunn, S. (2020). *Technology-facilitated gender-based violence: An overview* (Supporting a Safer Internet Paper No. 1). Centre for International Governance Innovation. https://www.cigionline.org/publications/technology-facilitated-gender-based-violence-overview/

European Union. (2012). *Charter of Fundamental Rights of the European Union*. *Official Journal of the European Union, C 326*, 391–407.

Fernandez, A. (2021). "Deep fakes": Disentangling terms in the proposed EU Artificial Intelligence Act. *UFITA, 85*(2), 392–433. https://doi.org/10.5771/2568-9185-2021-2-392

Fox, N. (2025). *Protecting human dignity and meaningful work in the age of AI: A social autonomy framework for ethical decision making in the workplace* [Honors thesis, Bates College]. SCARAB. https://scarab.bates.edu/honorstheses/496

Gellers, J. C., & Gunkel, D. (2022). Artificial intelligence and international human rights law: Implications for humans and technology in the 21st century and beyond. In A. Zwitter & O. J. Gstrein (Eds.), *Handbook on the politics and governance of big data and artificial intelligence*. SSRN. https://ssrn.com/abstract=4072896

Goodyear, M. (2025). Dignity and deepfakes. *Arizona State Law Journal, 57*, 931. https://doi.org/10.2139/ssrn.5199514

Gupta, K. (2023). The future of deepfakes: Need for regulation. *NLUD Journal of Legal Studies, 5*, 130–162.

Hawkins, W., Mittelstadt, B., & Russell, C. (2025). Deepfakes on demand: The rise of accessible non-consensual deepfake image generators. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1602–1614). Association for Computing Machinery. https://doi.org/10.1145/3715275.3732107

Heugas, A. C. (2021). Protecting image rights in the face of digitalization: A United States and European analysis. *The Journal of World Intellectual Property, 24*, 344–367. https://doi.org/10.1111/jwip.12194

Ibrahim, N. T., & Attia, N. A. (2026). The impact of disinformation generated by AI on democracy case studies: The US presidential elections in 2016 & 2024. *Review of*

Economics and Political Science, 11(3), 186–201. https://doi.org/10.1108/REPS-12-2024-0104

Innerarity, D. (2024). *Artificial intelligence and democracy*. UNESCO. https://hdl.handle.net/1814/77333

Iturmendi Rubia, J. M. (2024). Inteligencia artificial y derechos humanos: Desafíos y oportunidades en la era digital. *Deusto Journal of Human Rights, 14*, 11–31.

Jimoh, M. (2023). A prolegomenon on deepfakes and human rights in the African Charter. *Global Campus Human Rights Journal, 7*, 49–66. https://doi.org/10.25330/2663

Jurcys, P., Kozuka, S., & Fenwick, M. (2026). *A private law framework for personal AI replicas: A comprehensive analysis of legal issues*. SSRN. https://ssrn.com/abstract=6500358

Justo-Hanani, R. (2026). Risk-based approach to EU AI Act: Benefits and challenges of co-regulation. *Policy Design and Practice, 9*(2), 171–180. https://doi.org/10.1080/25741292.2025.2610869

Kadioglu Kumtepe, C. C., & Riley, S. (2024). Digital dignity: The shibboleth of digitalization in Europe? *Law, Technology and Humans, 6*(3), 156–169. https://doi.org/10.5204/lthj.3547

Leibowicz, C. R. (2026). Regulating reality: Exploring synthetic media through multistakeholder AI governance. *Policy & Internet, 18*, Article e70049. https://doi.org/10.1002/poi3.70049

Li, Q., Santo, W. L., Schoenebeck, S., & Gilbert, E. (2026). Position: AI/ML deepfake research is misaligned with AI-generated non-consensual intimate imagery (AIG-NCII). *arXiv*. https://doi.org/10.48550/arXiv.2607.18263

Li, Q., Zhang, S., Kasper, A. T., Ashkinaze, J., Eaton, A. A., Schoenebeck, S., & Gilbert, E. (2024). Reporting non-consensual intimate media: An audit study of deepfakes. *arXiv*. https://doi.org/10.48550/arXiv.2409.12138

Łabuz, M. (2025). A teleological interpretation of the definition of deepfakes in the EU Artificial Intelligence Act—A purpose-based approach to potential problems with the word "existing." *Policy & Internet, 17*, e435. https://doi.org/10.1002/poi3.435

Makauskaite-Samuole, G. (2025). Transparency in the labyrinths of the EU AI Act: Smart or disbalanced? *Access to Justice in Eastern Europe, 8*, 38–68. https://doi.org/10.33327/AJEE-18-8.2-a000105

Meding, K., & Sorge, C. (2025). What constitutes a deep fake? The blurry line between legitimate processing and manipulation under the EU AI Act. In *Proceedings of the 2025 Symposium on Computer Science and Law* (pp. 152–159). Association for Computing Machinery. https://doi.org/10.1145/3709025.3712218

Michałkiewicz-Kądziela, E., & Staszewski, W. S. (2026). Right to be forgotten in the deepfake era. *Teka Komisji Prawniczej PAN Oddział w Lublinie, 19*(1), 153–165. https://doi.org/10.32084/tkp.10696

Mishra, A., & Dash, B. (2026). Deepfakes and the crisis of digital consent: Artificial intelligence and synthetic media violations against women in Asia. *Media, Culture & Society, 0*(0).

Mochizuki, Y., Bruillard, E., & Bryan, A. (2025). The ethics of AI or techno-solutionism? UNESCO's policy guidance on AI in education. *British Journal of Sociology of Education*, 1–22. https://doi.org/10.1080/01425692.2025.2502808

Nasution, A. V. A., Suteki, & Lumbanraja, A. D. (2025). Addressing deepfake pornography and the right to be forgotten in Indonesia: Legal challenges in the era of AI-driven sexual abuse. *International Journal for the Semiotics of Law, 38*, 2489–2517. https://doi.org/10.1007/s11196-025-10265-0

Nowak, A., Tóth, B., & Ionescu, A. (2023). Legal challenges of deepfakes: Liability, harm, and regulatory responses. *Legal Studies in Digital Age, 2*(2), 49–60. https://jlsda.com/index.php/lsda/article/view/306

Organization of African Unity. (1981). *African Charter on Human and Peoples' Rights*.

Organization of American States. (1969). *American Convention on Human Rights*.

Pathak, S. (2026). *Private law challenges of AI-generated content in cross-border digital markets*. SSRN. https://ssrn.com/abstract=6241238

Patil, S. S., & Mishra, S. P. (2026). Platform liability and deepfake pornography: Are India's intermediary rules fit for the AI age? *LawFoyer International Journal of Doctrinal Legal Research, 4*(1), 1146–1177. https://doi.org/10.70183/lijdlr.2026.v04.50

Perriello, L. E. (2026). Blurred realities: Legal strategies for the deepfake era. *Maastricht Journal of European and Comparative Law, 0*(0).

Pleskach, M. (2025). Legal loopholes in the EU Artificial Intelligence Act: Addressing disinformation and human rights protection—the need for further refinement. In *2025*

15th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 852–858). IEEE. https://doi.org/10.1109/ACIT65614.2025.11185881

Poibeau, T. (in press). *Understanding conversational AI*. Ubiquity Press.

Poovarsan, G., Deenadhayalan, P., Manikkam, K., Ks, S., Rajesh, S., & Jeeva, V. (2026). Digital doppelgangers and data theft: A human-centric security analysis of AI and avatars. In *2026 International Conference on AI-Driven Innovations & Applications (IC-AIDA)* (pp. 1–6). IEEE. https://doi.org/10.1109/IC-AIDA68291.2026.11564210

Prasad, N. L., & Jain, I. D. (2025). Technological progress vs. human rights: Privacy, free expression, and reputation in the era of AI and deepfakes. *Journal of Legal Research and Polity, 2*(2), 89–102. https://doi.org/10.64322/JLRP.2025.2207

Quick, J. (2026). *Extraterritoriality in the age of cyber conflict: Jurisdictional doctrine, functional nexus, and the national security imperative*. SSRN. https://ssrn.com/abstract=6382178

Ramos, G. (2022). A.I.'s impact on jobs, skills, and the future of work: The UNESCO perspective on key policy issues and the ethical debate. *New England Journal of Public Policy, 34*(1), Article 3. https://scholarworks.umb.edu/nejpp/vol34/iss1/3

Rayun, S. M. N., Salam, M. A., Leong, V. S., Islam, W., Sahabuddin, M., Ahmed, M., Hosen, M. S., & Das, C. (2026). Deepfakes and digital sovereignty: Governance challenges of synthetic media in the AI-driven data economy. In M. Anshari, M. N. Almunawar, & H. Kifle (Eds.), *Digital sovereignty in the global data economy: Rights, regulation, and resistance* (pp. 225–264). IGI Global. https://doi.org/10.4018/979-8-2600-0288-9.ch008

Regulatory challenges in synthetic media governance: Policy frameworks for AI-generated content across image, video, and social platforms. (2024). *Journal of Robotic Process Automation, AI Integration, and Workflow Optimization, 9*(12), 36–54. https://helexscience.com/index.php/JRPAAIW/article/view/2024-12-13

Rijsbosch, B., van Dijck, G., & Kollnig, K. (2025). Adoption of watermarking for generative AI systems in practice and implications under the new EU AI Act. *arXiv*. https://doi.org/10.48550/arXiv.2503.18156

Rockwell, C., & Conklin, M. (2026). *Synthetic authority: Ghost influencers, deepfakes, and the structural limits of disclosure-based financial regulation*. SSRN. https://ssrn.com/abstract=6815680

Romero Moreno, F. (2024). Generative AI and deepfakes: A human rights approach to tackling harmful content. *International Review of Law, Computers & Technology, 38*(3), 297–326. https://doi.org/10.1080/13600869.2024.2324540

Saenz, N. (2024). Let's be realistic: Crafting an effective legal remedy for victims of deepfake pornography. *Arizona Law Review, 66*(3), 785–813. https://journals.librarypublishing.arizona.edu/arizlrev/article/id/8394/

Santomà, I. (2026). *Tools of harm: AI, deepfake infrastructure, and corporate liability after the Almendralejo scandal*. SSRN. https://ssrn.com/abstract=6500702

Saxena, P., & Pathak, G. (2025). Deepfakes, big tech, and human rights challenges: Examining the technology from a feminist legal lens. *The International Journal of Human Rights*, 1–28. https://doi.org/10.1080/13642987.2025.2602137

Scharffs, B. G., & Brown, A. (2025). Human dignity as a tool for differentiating between maleficent and beneficent uses of artificial intelligence. *The Review of Faith & International Affairs, 23*(3), 5–15. https://doi.org/10.1080/15570274.2025.2531652

Synodinou, T. (2014). Image right and copyright law in Europe: Divergences and convergences. *Laws, 3*(2), 181–207. https://doi.org/10.3390/laws3020181

Taeihagh, A. (2025). Governance of generative AI. *Policy and Society, 44*(1), 1–22. https://doi.org/10.1093/polsoc/puaf001

Ul Islam, M., & Fang, W. (2026). From prompts to motion: Diffusion transformers, synthetic media governance, and the future of AI-generated video generation models—A critical analysis of AI video generation technologies, architectures, applications, and future trajectories. *SSRN*. https://doi.org/10.2139/ssrn.6651939

Ulgen, O. (2017). Kantian ethics in the age of artificial intelligence and robotics. *Questions of International Law, 43*, 59–83.

Umbach, R., Henry, N., Beard, G. F., & Berryessa, C. M. (2024). Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 779, pp. 1–20). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642382

United Nations General Assembly. (1948). *Universal Declaration of Human Rights*.

United Nations General Assembly. (1966). *International Covenant on Civil and Political Rights*.

Viola, M., & Voto, C. (2023). Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. *Synthese, 201*, Article 30. https://doi.org/10.1007/s11229-022-04012-2

Williams, K. (2025). “There are no limits!”: AI undressing apps and the normalization of nonconsensual intimate deepfakes. *Violence Against Women, 0*(0).