UNIVERSITÄT GRAZ

European Master's Programme in Human Rights and Democratisation

A.Y. 2023/2024

A Modern Tale of Frankenstein?;

How to regulate non-consensual sexually explicit AI-generated deepfakes in the Metaverse

Author: Elise Kolen

Supervisors: Univ.-Prof. Mag. Dr.iur. Wolfgang Benedek & Ambassador Prof. Dr. Helmut Tichy

Word Count Declaration: 29.161

Abstract

This thesis examines the emerging issue of non-consensual sexually explicit deepfakes (sexfakes) in the context of the Metaverse. The research focuses on two key issues: sexfakes as a gender-based phenomenon, and the applicability of current regulations on sexfakes in the Metaverse. The thesis takes a euro-centric approach and therefore analyses the current regulatory frameworks established by the Council of Europe (CoE) and the European Union (EU) on sexfakes and the Metaverse. The study reveals that while progress has been made in addressing sexfakes, current protections remain incomplete, leaving women inadequately safeguarded. The Metaverse presents further unique challenges in the regulation and enforcement of sexfakes, raising complex questions about identity, legal personality, avatar ownership, harm, perpetrator accountability, jurisdiction, and the right to be forgotten. This thesis argues that the EU's and CoE's stance that no new instruments are necessary may underestimate the transformative potential of the Metaverse, especially concerning gender-based cyber violence. It proposes to enhance current frameworks to specifically address sexfakes in the Metaverse by taking a proactive regulatory stance so that the Metaverse could be transformed into a safe and inclusive digital space for all.

Keywords: artificial intelligence, gender-based violence, deepfakes, non-consensual sexually explicit deepfakes, Metaverse, European Union, Council of Europe.

Acknowledgements

The journey of completing this thesis has been both challenging and rewarding and has been made possible by the support and contributions of multiple individuals and institutions.

First and foremost, I extend my heartfelt gratitude to my supervisors, Professor Wolfgang Benedek and Ambassador Professor Helmut Tichy. Their unwavering guidance, insightful feedback, and profound expertise have been instrumental in shaping this work.

Second, I would like to thank the University of Graz for hosting me in the second semester of the EMA programme. A special thanks goes to Mag. Elias Faller and Professor Gerd Oberleitner for their warm welcome at the university and making me feel at home in Graz.

I am also deeply indebted to several individuals who generously shared their time and knowledge, providing valuable perspectives on the topic of my thesis. Particular thanks go to Carlotta Rigotti PhD (Post-doc researcher at eLaw – Center for Law and Digital Technologies, Leiden University), whose academic insights challenged and refined my thinking; Nuno Garcia, Phd (Associate Professor with Habilitation at the Computer Science Department at UBI and Invited Associate Professor at the Universidade Lusófona de Humanidades e Tecnologias), a computer scientist whose technical knowledge helped bridge the gap between technology and human rights in my research; and Elif Sariaydin (Administrator at the GREVIO Secretariat), whose expertise in gender-based violence and the work of GREVIO was invaluable.

The opportunity to participate in the Euregio Summer School on AI and Human Rights in Toblach Dobaccio, Italy, and the Global Conference on AI and Human Rights hosted by the University of Ljubljana were valuable experiences in broadening my perspective on the topic as well. These experiences brought my thesis to life and made me enthusiastic to pursue a further career in this field.

Lastly, I would like to acknowledge the unwavering support of my family, friends, and fellow EMA colleagues who have been a constant source of encouragement throughout this academic journey. Their belief in me and this project has been a driving force behind its completion.

Table of Contents:

List of Abbreviations **6**

Chapter 1: Introduction **7**

Chapter 2: Sexfakes in the Metaverse **14**

2.1. Introduction *14*

2.2. What are Deepfakes? *14*

2.3. Advantages and Risks of Deepfakes *15*

2.4. Sexfakes *17*
2.4.1. What are Sexfakes? 17
2.4.2. Cyber Sextortion 18
2.4.3. Why this is a Gendered Issue 19
2.4.4. Effects on Female Victims 22

2.5. A New Threat Ahead: the Metaverse *25*
2.5.1. The Metaverse Explained 25
2.5.2. Sexfakes in the Metaverse 30

2.6. Conclusion *30*

Chapter 3: Regulatory Framework on Deepfakes **32**

3.1. Introduction *32*

3.2. Regulatory Framework of the Council of Europe on Sexfakes *32*
3.2.1. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law 32
3.2.2. Convention on Cybercrime 34
3.2.3. European Convention on Human Rights 35
3.2.4. Istanbul Convention & GREVIO General Recommendation No. 1 42

3.3. Legal Framework of the European Union on Sexfakes *46*
3.3.1. General Data Protection Regulation 47
3.3.2. Audiovisual Media Services Directive 51
3.3.3. Digital Services Act 52
3.3.4. Artificial Intelligence Act 53
3.3.5. Directive on Combating Violence against Women and Domestic Violence 56

3.4. Conclusion *60*

Chapter 4: Regulating Sexfakes in the Metaverse **63**

4.1. Introduction *63*

4.2. Online World vs Metaverse: Same System in a Different Font? *63*
4.2.1. Identity 63
4.2.2. (Legal) Personality 65
4.2.3. Ownership 66
4.2.4. Harm 68
4.2.5. Perpetrator 70
4.2.6. Jurisdiction and Statehood 71
4.2.7. Right to be Forgotten 72

4.3. Who Should Regulate the Metaverse? *74*
4.3.1. Self-Regulation 74
4.3.2. Regulation 77

4.4. A Broader Perspective: Solutions Outside of the Regulatory Framework *82*

4.5. Conclusion *84*

5. General Conclusions **85**

Bibliography **88**

Annex **110**

List of Abbreviations

AIA Artificial Intelligence Act

AR Augmented Reality

AVMSD Audio Visual Media Services Directive

CJEU Court of Justice of the European Union

CFR Charter of Fundamental Rights of the European Union

DSA Digital Service Act

EU European Union

EC European Commission

ECtHR European Court of Human Rights

ECHR European Convention on Human Rights

EP European Parliament

FGM female genital mutilation

GAN Generative Adversarial Networks

GDPR General Data Protection Regulation

GREVIO Group of Experts on Action against Violence against Women and Domestic Violence

IC Istanbul Convention

Member States MS

OJ Official Journal

para. paragraph

p. page number

T-CY Cybercrime Convention Committee

TFEU Treaty on the Functioning of the European Union

TEU Treaty on European Union

UN United Nations

VR Virtual reality

Chapter 1: Introduction

In January 2024, the famous singer Taylor Swift made headlines when she was one of the latest victims of sexually explicit non-consensual deepfakes (also known as sexfakes).[1] Malicious netizens spread the fake sexual content on social media platforms like X (formerly known as Twitter).[2] Even though Taylor Swift's fans (called Swifties) quickly came to her rescue by reporting the violative content, this case is only one of the many examples in which non-consensual sexfakes have been distributed on the internet, affecting both famous and non-famous women.[3] The rapid developments of AI and the widespread availability and facilitated use of tools to create deepfakes will make it even easier to create and disseminate sexfakes in the future, going from a high impact, low likelihood risk technology, to a high impact, high likelihood risk technology.[4] Unfortunately, this technology will mostly impact women, expanding the widespread gender-based violence and discrimination against women – which has been affecting society for a long time – from the real world to the digital sphere.[5]

An additional blurring of the lines between real and fake content will be the introduction of the Metaverse. Tech giants like Mark Zuckerberg are introducing platforms like the Facebook Metaverse 'Meta'[6] which is a universal virtual world that offers users an immersive experience akin to reality, creating a seamless blend between physical and virtual realms.[7] IT consultancy firm Gartner predicts that by 2026, a quarter of the population will spend at least an hour daily in the Metaverse, engaging with among others, colleagues, teachers, retailers and friends.[8]

---

[1] Contreras, B (2024), 'Tougher AI Policies Could Protect Taylor Swift—And Everyone Else—From Deepfakes'. Retrieved from: https://www.scientificamerican.com/article/tougher-ai-policies-could-protect-taylor-swift-and-everyone-else-from-deepfakes/ 4 March 2024.
[2] Ibid.
[3] Saner, E. (2024), 'Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'. Retrieved from: https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box+on+4+March+2024 on 4 March 2024; RTL Nieuws, 'Werkstraf van 120 uur geëist voor deepfakevideo Welmoed Sijtsma' retrieved from: https://www.rtl.nl/nieuws/artikel/5414062/werkstraf-geeist-voor-deepfake-welmoed-sijtsma on 4 March 2024.
[4] Huijstee, M. V. et al. (2021). p. 16.
[5] Laffier, J. & Rehman, A. (2023).; Almenar, R. (2021).
[6] Milmo, D. (2021), 'Enter the Metaverse: the digital future Mark Zuckerberg is steering us toward' retrieved from: https://www.theguardian.com/technology/2021/oct/28/facebook-mark-zuckerberg-meta-Metaverse on 4 March 2024.
[7] Europol (2022)., p. 7-8.
[8] Gartner (2024), 'Gartner Predicts 25% of People Will Spend At Least One Hour Per Day in the Metaverse by 2026'. Retrieved from: https://www.gartner.com/en/newsroom/press-releases/2022-02-07-gartner-predicts-25-percent-of-people-will-spend-at-least-one-hour-per-day-in-the-metaverse-by-2026 on 9 July 2024; Woollacott, E. (2022). 'Rise of deepfakes: who can you trust in the Metaverse?' retrieved from https://cybernews.com/security/rise-of-deepfakes/ on 4 March 2024.

Unfortunately, we will also see a rise in deepfakes in the Metaverse.[9] Deepfakes can easily be made by reproducing – or stealing – another person's virtual identity, making it indistinguishable from the original.[10] These compromised identities could then be exploited to create non-consensual sexfakes like those currently online.[11] The creation and dissemination of sexfakes is, therefore, a current problem that will also stay relevant when more and more people are going to spend their time in the Metaverse in the future.

Regulating sexfakes feels like a modern-day retelling of Frankenstein.[12] Just as Mary Shelley's protagonist created an artificial being that ultimately escaped his control, deepfakes, while having the potential for positive applications, have now grown completely out of hand with the extensive creation of sexfakes.
While Frankenstein's monster was crafted from human remains, our 'digital monsters' are forged from our own easily surrendered data. As I am examining the CoE's and EU's regulatory framework on sexfakes in the Metaverse, the same fundamental questions raised by Shelley's book are relevant to my thesis: What are the ethical limits of technological innovation? What responsibilities do creators have towards their creations? And maybe most importantly, how can we control and even stop our creations once they have been released into the world?

Relevance of the thesis

Even though academia and legislators have been paying more and more attention to the impacts of deepfakes by warning against the wide dissemination of fake news through deepfakes and interference in political elections, the impact of non-consensually sexually explicit deepfake content on victims and the gendered aspect of the offence is currently understudied. Furthermore, no sufficient link has been made yet with the Metaverse. It is imperative to also consider future challenges and possible regulations in order to prevent legislators from always being one step behind. With my thesis, I am aiming to make a first step in the right direction. That is why I shall research the current regulatory framework on sexfakes and assess how future-proof these regulations are in the Metaverse.

[9] Europol (2022)., p. 13.
[10] Ibid.
[11] Woollacott, E. 'Rise of deepfakes: who can you trust in the Metaverse?' retrieved from https://cybernews.com/security/rise-of-deepfakes/ on 4 March 2024.
[12] Shelley, M. (2012).

Research question & sub-questions:

Research question 1: *To what extent are women*[13] *protected under the European Union and Council of Europe framework in cases of non-consensual sexually explicit deepfake content?*

Research question 2: *To what extent is the current regulatory framework on non-consensual sexually explicit deepfake content applicable in the Metaverse?*

Sub-questions:

- What are deepfakes and sexfakes?
- How are sexfakes related to gender-based cyber violence?
- What is the Metaverse and how is it related to sexfakes?
- What is the existing CoE and EU regulatory framework on deepfakes?
- Is the current EU and CoE regulatory framework on deepfakes applicable in the Metaverse?

Methodology

This thesis will consist of desk research, basing its findings on an analysis of scientific and grey literature, as well as relevant policies. My research includes a review of scholarly literature, reports from the EU and CoE institutions, legal blog posts and media and news articles, which will provide me with an insight into how sexfakes are, and should be, regulated in the Metaverse.

The searches for this literature were conducted through the use of various databases like J-STOR, Google Scholar, Elsevier, and SSRN. I filtered the results according to the 'inclusion' and 'exclusion' criteria.[14] Inclusion criteria were: sources written in English, Dutch and German, publications from 2015 onwards, literature focused on the relation between deepfakes and gender-based violence and literature on the Metaverse. Exclusion criteria were: articles not written in English, Dutch and German, sources written before 2015, literature not focussing on

[13] I am using the term "women" in the broadest sense of the word, including persons assigned a female sex at birth and persons who define themselves as a woman.
[14] Popay, J., et al. (2006). p. 92.

the relation between deepfakes and gender-based violence, literature on the Metaverse, and sources that had restricted access and could not be retrieved by me.[15] On top of this, literature was manually gathered by searching the citations in the previously identified articles and reading articles that were recommended to me by my supervisors and other relevant scholars in the field.

Key words used in searching for literature were: non-consensual sexually explicit deepfakes, gender-based (cyber) violence, image-based sexual abuse, deepfake pornography, sexfakes, cyber (s)extortion, violence against women, international human rights framework, Metaverse, virtual world, avatar, EU legislation, Council of Europe Conventions. By employing Boolean operators to combine the aforementioned terms, I was able to compile an extensive list of articles that discussed the European framework on sexfakes in the Metaverse.

This thesis takes a euro-centric approach to find solutions for the issue of non-consensual sexfakes, focusing on the EU and CoE fundamental/human rights framework. I am aware that the UN is also taking relevant steps on this topic.[16] For example, UNESCO has adopted its Recommendation on the Ethics of Artificial Intelligence.[17] However, due to the word count of this thesis, I have decided not to include the UN framework in my research. Furthermore, Europe (and especially the EU) wants to be one of the key players in regulating new technologies.[18] The CoE and EU regulations could therefore serve as a potential blueprint for other (international) jurisdictions.[19]

The EU and CoE legal framework will be analysed from a legal doctrinal point of view. The main components and key regulations will be identified, analysed and synthesised to offer a comprehensive commentary on the substance. This approach aims to capture the advancements that have already been achieved over the years in regulating deepfakes.

---

[15] Laffier, J. & Rehman, A. (2023)., p. 3-4.
[16] See for example: United Nations Human Rights Council (UNHRC) Resolution 53/27 rev.1 (New and emerging digital technologies and human rights, adopted at the 53rd Regular Session (19 June – 14 July 2023); Report of the Human Rights Council Advisory Committee on the Possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights, adopted by the United Nations Human Rights Council (UNHCR) at the 47th session (21 June–9 July 2021); Report of the Human Rights Council Advisory Committee on the Impact of new technologies intended for climate protection on the enjoyment of human rights, adopted by the United Nations Human Rights Council (UNHCR) at the 54th session (11 September–6 October 2023).
[17] UNESCO, Recommendation on the Ethics of Artificial Intelligence. Adopted on 23 November 202.
[18] Dolan, L. (2022). p. 8.
[19] Ibid.

In the EU context, I will look at, the General Data Protection Regulation (GDPR)[20], the Audio Visual Media Services Directive (AVMSD)[21], the Digital Service Act (DSA)[22], the Artificial Intelligence Act (AIA)[23], and the Directive on Combating Violence against Women and Domestic Violence (2024/1385)[24]. The first three are relevant as they include data protection measures; the AIA is significant for its transparency obligation regarding deepfakes; and Directive 2024/1385 recognises sexfakes as a form of gender-based violence.
In the CoE context, I will look at the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law[25], the Cybercrime Convention[26], Articles 3 (degrading treatment), 8 (privacy and physical integrity) and 10 (freedom of expression) ECHR[27], the Istanbul Convention (IC)[28] and the GREVIO General Recommendation No. 1. on the Digital Dimension of Violence against Women[29].

The examined regulations represent the current framework. However, in my research, I will also assess the risks of deepfakes in the context of the Metaverse. Currently, the Metaverse is neither legally defined nor regulated[30], which means that we have to analyse in what way the current legal framework will also be applicable in the future in the Metaverse. In my thesis, I

[20] European Parliament and Council, Regulation 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016.
[21] European Parliament and Council, Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive, OJ L 303, 28.11.2018.
[22] European Parliament and of the Council, Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277, 27.10.2022.
[23] European Parliament and of the Council, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance, OJ L, 2024/1689, 12.7.2024.
[24] European Parliament and Council, Directive (EU) 2024/1385 of the European Parliament and of the Council of 14 May 2024 on combating violence against women and domestic violence, OJ L, 2024/1385, 24.5.2024.
[25] Council of Europe (2024). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. CETS No. 224.
[26] Council of Europe (2001). Convention on Cybercrime (Budapest Convention). ETS No. 185.
[27] Council of Europe (1950). European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14. ETS No. 161.
[28] Council of Europe Convention on preventing and combating violence against women and domestic violence (Istanbul Convention) (2011). CETS No. 210.
[29] Group of Experts on Action against Violence against Women and Domestic Violence (GREVIO) (2021). General Recommendation No. 1 on the digital dimension of violence against women. Council of Europe Publication Office.
[30] Pellegrini, C. (2023). 'Conflict of Laws and the Metaverse'. Retrieved from: https://eapil.org/2023/06/13/conflict-of-laws-and-the-Metaverse/ on 4 March 2024.

will pinpoint the specific challenges concerning regulation in the Metaverse and analyse whether the current framework suffices or whether we have to find new innovative ways to regulate the Metaverse.

Disclaimer: I would like to end this methodology section with a disclaimer. Given the recent rise of questions related to sexfakes in the Metaverse, there is a chance that I may have missed some of the relevant literature that could have met the study's selection criteria.[31] During the writing of my thesis, new research and even new regulations have been adopted on both the EU and CoE level (with the AIA being published in the OJ on 12 July 2024, one day before the submission of this thesis). I have tried to keep up with all of these new developments to the best as I can, but there is a chance I overlooked some developments. Despite these constraints, I am confident that this thesis effectively reflects the existing literature and regulatory landscape concerning sexfakes in the Metaverse.[32]

Structure

In Chapter 2, I will provide a comprehensive overview of deepfakes, starting with a general explanation of what deepfakes are, including both their benefits and potential risks. Following this, I will delve into the specific case study of my thesis: sexfakes. This section will detail what sexfakes are and how they disproportionately target women. I will then introduce an additional dimension to the topic by examining the rise of sexfakes within the Metaverse.

In Chapter 3, the discussion will shift to the regulatory frameworks of both the EU and CoE on deepfakes. Most analyses frame the issue of deepfakes as a privacy problem. In this chapter, however, I will argue that sexfakes are not only a breach of privacy but also a form of gender-based violence and should therefore be regulated accordingly.

In Chapter 4, I will assess whether the current regulatory framework is applicable in the Metaverse. This chapter will explore the critical differences between the Metaverse with the online world and conclude that we need a reevaluation of the existing regulatory structures. Afterwards, I will investigate who should be responsible for regulating sexfakes in the Metaverse; is self-regulation by Metaverse platform providers enough or should the CoE and

[31] Laffier, J. & Rehman, A. (2023)., p. 3-4.
[32] Ibid.

EU step in, and how? At the end of Chapter 4, I will end with taking a broader perspective on tackling sexfakes in the Metaverse outside of the regulatory framework.

Lastly, I will end my thesis with a general conclusion.

Chapter 2: Sexfakes in the Metaverse

2.1. Introduction

In this chapter, a general explanation of deepfakes will be provided, while highlighting both their advantages and potential dangers. Following that, I shall focus on the main subject of my thesis: sexfakes. I shall define sexfakes and discuss how they disproportionately affect women. Afterwards, I will explore an additional aspect of my topic, which is the use of sexfakes within the Metaverse. Finally, a conclusion of the chapter will be drawn with a summary of my findings.

2.2. What are Deepfakes?

According to the Merriam-Webster dictionary, a deepfake is: "an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said".[33] Through machine learning-based software and deep learning, realistic content is made in which an individual's likeness is replaced with another.[34] The primary used machine-learning technology is Generative Adversarial Networks (GAN), which has improved the quality and resolution of produced deepfakes, while requiring minimal time and cost investment.[35] For example, by processing a thousand photos of Donald Trump, GAN can generate a new photo of Trump which is not an exact replica of a single source, but does look genuine.[36] The same goes for video material. With the same technology, we can make Donald Trump (or any other person) say or do things that they have never said or done before.[37]

Previously, the creation of deepfakes demanded significant computing power as well as programming expertise.[38] Nowadays, however, instead of needing hundreds of images of a

[33] Definition deepfake. Retrieved from: https://www.merriam-webster.com/dictionary/deepfake on 8 April 2024.
[34] Laffier, J. & Rehman, A. (2023). p. 2.
[35] Sloot, B. et al. (2021). p. 2.
[36] Ibid.
[37] Ibid.
[38] Ibid.

person to create a believable deepfake, one or two are enough.[39] Furthermore, there are easily accessible deepfake applications made for mobile phones, making the process more user-friendly overall.[40] It is not even necessary for the producer of a deepfake video to possess images or videos that resemble the intended final product: some apps permit users to produce nude images of individuals who in their original photo are fully clothed.[41]

Consumer-friendly applications have already made their way into the mainstream consumer market. For example, the website 'geeksforgeeks'[42] or 'masoative'[43] have articles ranking the 'Top Deepfake Generating Apps', comparing different applications and listing the pros and cons of every deepfake software program. Unsurprisingly, the creation of deepfakes has skyrocketed over the last few years, growing by 550% from 2019 to 2023 to a total number of 95,820 deepfakes on the internet.[44]

2.3. Advantages and Risks of Deepfakes

Even though the term 'fake' has a negative connotation to it, deepfakes can be used in a positive manner.

For example, in the film industry deepfakes have been used in the past for various blockbusters. After the famous actor Paul Walker died in a car crash in 2013, he still played a part in a sequel of the film series 'Fast and Furious'.[45] Deepfake technology was used to copy Paul Walker's face onto his brothers, who acted out the scenes in real life to create an opportunity for the other actors to interact with 'Paul Walker'.[46]

---

[39] Sloot, B. et al. (2021). p. 2.
[40] Ibid.
[41] For example, the website 'ClothOff' (I will not cite the full website link in my research as I find it unethical to refer to websites that are making these type of content); Sloot, B. et al. (2021). p. 2.
[42] See for the website of geeksforgeeks: https://www.geeksforgeeks.org/top-5-deepfake-generating-apps/
[43] See for the website of masoative: https://www.masoative.com/post/best-deepfake-apps
[44] Home security heroes (2023), 'State of Deepfakes'. Retrieved from: https://www.homesecurityheroes.com/State-of-deepfakes/ on 13 March 2024.; Krishnan, M. (2023), 'Can India tackle deepfakes?' retrieved from: https://www.dw.com/en/can-india-tackle-deepfakes/a-67791106 on 9 April 2024.
[45] Faciaai, 'How did Paul Walker appear in Fast 7 after his death?' retrieved from: https://faciaai.medium.com/how-did-paul-walker-appear-in-fast-7-after-his-death-fda6acfea096 on 13 March 2024.
[46] Okolie, C. (2023). p. 6.

Deepfakes can also be used to help generate more empathy from the general public for victims of far-away disasters.[47] UNICEF and MIT have been working together on a project called 'Deep Empathy', in which AI is employed to generate digital replicas and deepfakes of the viewer's home location under comparable disaster conditions as those of victims living far away.[48] The goal is to bring the viewer closer to the experiences of those most affected, creating more understanding and compassion for victims.[49]

Lastly, deepfakes can also be used by retailers and fashion brands who engage in e-commerce and advertising. Deepfake technology can be used to present fashion outfits on a diverse range of models, all varying in weight, height, skin tone, etc.[50] Deepfakes can even be used to transform customers into models themselves, using their face and body for virtual outfit fittings and previewing how a garment would look on them before buying.[51]

However, there are also risks associated with the creation and dissemination of deepfakes. As Okolie describes it, 'Deepfake technology is merely a tool, and like most tools, the ethics of its use will be dependent on who wields the handle'.[52] Consequently, deepfake technology has not only been used to benefit society but also to spread misinformation and hatred.

For example, in 2019, someone circulated a manipulated video of Nancy Pelosi (US Speaker of the House) on social media, making her appear intoxicated and speaking incoherently.[53] The widespread sharing of the video led to people being concerned for Pelosi's health and questioned her suitability for office.[54]
In India, Manoj Tiwari, the president of the Bharatiya Janata Party (BJP), disseminated a video of himself criticising the incumbent Delhi government of Arvind Kejriwal on WhatsApp.[55] This

[47] Okolie, C. (2023). p. 6.
[48] Ibid.
[49] See for the website of *Deep Empathy*: https://www.media.mit.edu/projects/deep-empathy/overview/.
[50] Westerlund, M. (2019). p. 41.
[51] Ibid.
[52] Okolie, C. (2023). p. 6.
[53] Sadiq, M. (2019), 'Real v fake: debunking the 'drunk' Nancy Pelosi footage – video'. Retrieved from: https://www.theguardian.com/us-news/video/2019/may/24/real-v-fake-debunking-the-drunk-nancy-pelosi-footage-video on 19 April 2024.
[54] Laffier, J. & Rehman, A. (2023)., p. 9.
[55] Jee, C (2020), 'An Indian politician is using deepfake technology to win new voters'. Retrieved from: https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/ on 8 April 2024; Nilesh, C. (2020).'We've Just Seen the First Use of Deepfakes in an Indian Election Campaign. Retrieved from: https://www.vice.com/en/Article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp on 8 April 2024.

video went viral, reaching approximately 15 million people.[56] Tiwari had such a big impact because, besides spreading his message in English, he also created a deepfake video of himself speaking Haryanvi, the Hindi dialect spoken by the large migrant worker population in Delhi who were mostly not voting for BJP at the time.[57] Tiwari hoped to reach this specific audience and dissuade them from voting for the rival political party.[58]

The widespread creation and dissemination of political deepfakes can have a grave impact on democratic processes around the world.[59] In times of elections, reputation is everything, and these so-called 'campaigns of distortion' could easily turn the tables in someone else's favour very quickly.[60] In the long run, misinformation created by deepfakes can lead to a general distrust of media outlets, governmental authorities, institutions, or citizens, creating social division.[61]

These examples are only the tip of the iceberg. Deepfakes could also be used for things like hoaxes, bullying, financial fraud, identity theft, extortion, etc.[62] However, the biggest problem that has arisen since the mass use of deepfakes in my opinion is the creation and dissemination of so-called 'sexfakes'.

2.4. Sexfakes

2.4.1. What are Sexfakes?

Most deepfakes involve copying the faces of non-consenting victims onto the bodies of another person, making them engage in sexual activities.[63] In fact, the first non-professionally made deepfake video was by a Reddit user, who posted pornographic deepfake videos of celebrities' faces pasted onto adult film stars' bodies.[64] The Reddit user used a free, open-source machine

[56] Jee, C (2020), 'An Indian politician is using deepfake technology to win new voters'. Retrieved from: https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/ on 8 April 2024; Nilesh, C. (2020).'We've Just Seen the First Use of Deepfakes in an Indian Election Campaign. Retrieved from: https://www.vice.com/en/Article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp on 8 April 2024.
[57] Ibid.
[58] Ibid.
[59] Okolie, C. (2023). p. 6.
[60] See: Openletter.net (2024), 'Deepfakes'. Retrieved from: https://openletter.net/l/disrupting-deepfakes on 22 June 2024.
[61] Okolie, C. (2023). p. 6-7.
[62] Ibid., p. 6.
[63] Ibid., p. 7.
[64] Laffier, J. & Rehman, A. (2023)., p. 2.

learning software to build the GANs necessary to make the deepfakes.[65] From that point onwards, non-consensual sexfakes took off. Today, these videos do not only depict celebrities but also average citizens. With the current technology, for example, any average person could generate a sexfake of a classmate/colleague/ex-partner/etc. and distribute it on social media platforms or to their relatives.[66] And if they are not that technically gifted, they can also commission someone else to do it: for only 30 Euros a deepfake creator can be commissioned to generate a personalised sexfake.[67] One of those websites is called 'ClothOff', which offers users to "undress anyone using AI".[68] The user uploads a photograph of a fully clothed person which the app then returns stripped of clothing.[69] The website of ClothOff has 4 million monthly visitors, despite being at the centre of two incidents in 2023 in which school girls in the U.S. and Spain were targeted with deepfakes by their fellow underaged classmates. ClothOff gained global news coverage because of these incidents, resulting in public outcry to ban the website, but it is still active to this day.[70]

To put things into perspective, 96% of all deepfake content is sexually explicit.[71] We have seen a rise of sexually explicit deepfake content of over 400% from 2022 to 2023, reaching 34 million monthly viewers in 2023.[72] These numbers show that the creation and dissemination of non-consensual sexfakes for profit is a lucrative business model at the expense of victims.

2.4.2. Cyber Sextortion

Another 'lucrative' business model tying in with sexfakes is the offence of 'cyber sextortion'. Cyber sextortion refers to the fact that a person threatens to share explicit sexual content online of a victim in order for them to comply with the abuser's specific demands (i.e. asking for more

[65] Laffier, J. & Rehman, A. (2023)., p. 2.
[66] Sloot, B. et al. (2021). p. 4.
[67] Compton, S. & Hamlyn, R. (2023), 'Opinion: The rise of deepfake pornography is devastating for women' retrieved from: https://edition.cnn.com/2023/10/29/opinions/deepfake-pornography-thriving-business-compton-hamlyn/index.html#:~:text=This%20practice%20is%20no%20longer,on%20victims%20can%20be%20devastating on 8 March 2024.
[68] Safi, A., Atack, A. (2024), 'Revealed: the names linked to ClothOff, the deepfake pornography app' retrieved from: https://www.theguardian.com/technology/2024/feb/29/clothoff-deepfake-ai-pornography-app-names-linked-revealed on 13 March 2024.
[69] Ibid.
[70] Ibid.
[71] Openletter.net (2024), 'Deepfakes'. Retrieved from: https://openletter.net/l/disrupting-deepfakes on 22 June 2024.
[72]Ibid.

sexual content, sending money, etc).[73] The main intent of this offence however is to gain control over the victim: the possible harm the abuser can inflict on the victim forms the crux of the offence.[74] Cyber sextortion differs from other forms of abuse as it is uncertain whether the videos/photos will truly be disseminated. Victims live in fear and desperation while being sextorted, not knowing if or when their images will be shared.[75]

Examples of cyber sextortion can be traced back to 2009[76] but deepfakes have changed the game once again. Where previously, cyber sextortion perpetrators befriended the victim by pretending to be someone else until they received sexual images/videos, deepfakes skip this step and are the means by which victims are held to ransom.[77] No 'real' content is needed anymore which means that everyone can be sextorted as long as they have a couple of (fully clothed) pictures of them circulating the internet. Even when viewers of the content know the footage is fake, social consequences for the victim can be grave.[78] Sexfakes have the same implications as fake news: even when a story is proven false later on, persons will stay convinced that, though the specific message may have been proven to be fake, the underlying truth remains valid.[79] Furthermore, the fake news will receive more attention than any subsequent retraction or correction as the message is often more sensational and attention-grabbing.[80] Persons who have read both the original article and the correction may still be left with a lingering feeling, thinking: "Wasn't there something wrong with...?"[81]

2.4.3. Why this is a Gendered Issue

Of the 96% of deepfake videos being of sexual nature, women and girls make up to 99% of the victims of these sexfakes, therefore making it an offence which almost exclusively targets women.[82] Deepfake applications are 'gendered by design' as the technology used to generate

[73] Laffier, J. & Rehman, A. (2023)., p. 8.
[74] Ibid.
[75] Ibid. p. 7.
[76] Wittes, B., et al. (2016). p. 1, 6-7.
[77] Muncaster, P. (2023), 'Deepfaking it: What to know about deepfake-driven sextortion schemes' retrieved from: https://www.welivesecurity.com/2023/07/04/deepfaking-it-deepfake-driven-sextortion-schemes/ on 8 April 2024.
[78] Sloot, B. et al. (2021). p. 4.
[79] Ibid.
[80] Ibid.
[81] Ibid.
[82] Tsalidis, A. (2024), 'Disrupting the Deepfake Pipeline in Europe' retrieved from: https://futureoflife.org/ai-policy/disrupting-the-deepfake-pipeline-in-europe/ on 9 April 2024; Openletter.net (2024), 'Deepfakes'. Retrieved from: https://openletter.net/l/disrupting-deepfakes on 22 June 2024; Compton, S. & Hamlyn, R. (2023), 'Opinion: The rise of deepfake pornography is devastating for women' retrieved from:

this content is predominantly tailored to the female body.[83] At the same time, detection models that are designed to distinguish real from synthetic content are mostly trained with data depicting the male body.[84] As a consequence, numerous victims may go months without realising that they have been targeted by a sexfake, making it possible for the content to be shared thousands, sometimes millions, of times in the meantime.[85] On the other hand, the perpetrators of this offence are predominantly men.[86]

The aforementioned statistics correlate to a pre-existing trend in cyber-related gender-based violence, in which 58% of young women and girls around the world have encountered online harassment on social media platforms.[87] Compared to men, women are three times more likely to face various forms of online abuse - including online harassment, threats, sexualisation, and stalking - and more than twice as likely to experience severe forms of abuse like doxing, revenge porn and threats of physical harm, compared to men.[88] When women identify themselves online as a woman, the risk of intentional abuse even increases.[89] Women working in the public or digital sphere face an even greater risk of abuse.[90]

Gender-based cyber violence cannot be understood without looking at the broader cultural and societal context of the offline world.[91] Offline misogynistic practices based on a view that women are inferior to men have evolved into the digital sphere.[92] When men feel threatened by their perceived loss of superior position, they will attempt to reaffirm the established power dynamics by inflicting abuse on women.[93] The emergence of sexfakes can be viewed as a result of the disregard for women and power dynamics, where sexfakes provide men with a new

---

https://edition.cnn.com/2023/10/29/opinions/deepfake-pornography-thriving-business-compton-hamlyn/index.html#:~:text=This%20practice%20is%20no%20longer,on%20victims%20can%20be%20devastating;; Tenbarge, K. 'Found through Google, bought with Visa and Mastercard: Inside the deepfake porn economy' retrieved from: https://www.nbcnews.com/tech/internet/deepfake-porn-ai-mr-deep-fake-economy-google-visa-mastercard-download-rcna75071 on 8 March 2024.

[83] [83] Huijstee, M.V. et al. (2021)., p. 9.

[84] Ibid.

[85] Tsalidis, A. (2024), 'Disrupting the Deepfake Pipeline in Europe' retrieved from: https://futureoflife.org/ai-policy/disrupting-the-deepfake-pipeline-in-europe/ on 9 April 2024; Openletter.net (2024), 'Deepfakes'. Retrieved from: https://openletter.net/l/disrupting-deepfakes on 22 June 2024.

[86] Rigotti, C. & McGlynn, C. (2022). p. 460; Laffier, J. & Rehman, A. (2023)., p. 6.

[87] Kolen, E. (2024), p. 9; Openletter.net (2024), 'Deepfakes'. Retrieved from: https://openletter.net/l/disrupting-deepfakes on 22 June 2024.

[88] Kolen, E. (2024), p. 9; Veletsianos, G., et al. (2018), p. 4691.

[89] Ibid.

[90] Ibid. Laffier, J. & Rehman, A. (2023), p. 6.

[91] Laffier, J. & Rehman, A. (2023)., p. 5.

[92] Ibid.

[93] Ibid.

method to assert dominance and control.[94] Sexfakes are used to compromise a woman's identity, tarnish her reputation, intimidate her, and compel silence in order to perpetuate and strengthen gender disparities.[95]

Furthermore, sexfakes bring to light the objectification of the female body in both the offline and online world.[96] The utilisation of a woman's face and body without her consent for a man's gratification assumes, perpetuates, and strengthens a virtual space where women's images are manipulated solely for the enjoyment of men who engage in these digital realms.[97] The non-consensual aspect of the sexfakes is crucial here as it forces women into sexual acts, thereby reducing them to mere objects of gratification, being able to be abused at whim.[98]

The bigger societal impact non-consensual sexfakes have on society is captured in the 'cascading effect of deepfakes' (see Figure 1). The consequences of a single sexfake go beyond a specific category of risk as it encompasses successive impacts at different levels. In the first place, as sexfakes typically affect individuals, the impact starts on an individual level.[99] The victim suffers psychological and reputational harm.[100] Subsequently, these sexfakes can also impact a specific group or organisation when they are distributed to for example colleagues or family of the victim.[101] Lastly, the notion of the existence of sexfakes in itself or the cumulative effect of sexfakes, may result in significant harm on a societal level. [102] Sexfakes may adversely affect sexual norms and morality in society.[103]

[94] Sloot, B. et al. (2021). p. 6.
[95] Laffier, J. & Rehman, A. (2023)., p. 5-6.
[96] Sloot, B. et al. (2021). p. 6.
[97] Van der Nagel, E. (2020). Verifying images: Deepfakes, control, and consent. *Porn Studies*, *7*(4), 424-429.
[98] Laffier, J. & Rehman, A. (2023)., p. 10.
[99] Huijstee, M.V. et al. (2021)., p. IV.
[100] Ibid.
[101] Ibid.
[102] Ibid.
[103] Ibid.

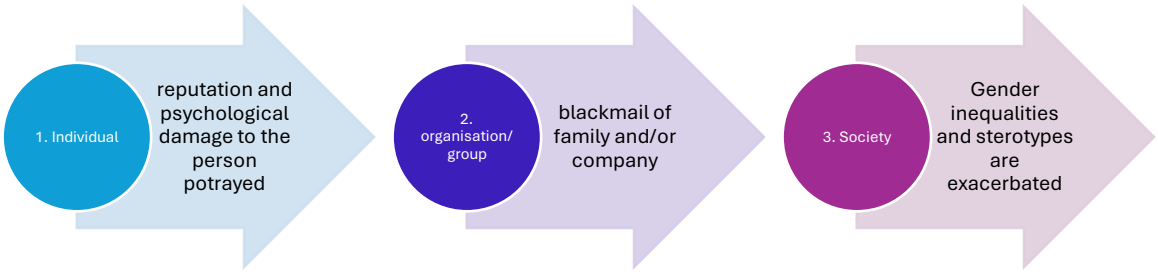


Figure 1: Cascading effect of pornographic deepfakes.
Graph inspired by: Huijstee, M.V. et al. (2021)., p. V.

2.4.4. Effects on Female Victims

The impact that sexfakes have on victims, is detrimental. Its ramifications extend to the professional, social and personal lives of victims due to fear of personal safety, moving physical locations and jobs, and losing friendships and family relations.[104] Victims experience strong feelings of fear, helplessness, powerlessness and humiliation.[105]

Especially for young girls, sexfakes can have a strong negative effect on their self-image: viewing themselves engage in various explicit actions can adversely impact their self-confidence and self-esteem.[106]

On top of that, some women and girls cannot talk with others about the reactions they receive due to shame and victim blaming (victims are often blamed for the generated content even though they never engaged in the acts shown).[107]

Lastly, more than 50% of victims are in the unknown of the identity of the perpetrator, intensifying their fear, distress, and sense of powerlessness.[108] Some victims' mental health deteriorates so much after the creation and dissemination of sexfakes that they experience suicidal thoughts.[109]

[104] Laffier, J. & Rehman, A. (2023)., p. 8.
[105] Molloy, S. (2019). 'Blackmailers-for-hire ae weaponising 'deepfake' revenge porn'. Retrieved from https://nypost.com/2019/01/02/blackmailers-for-hire-are-weaponizing-deepfake-revenge-porn/ on 8 April 2024.
[106] Sloot, B. et al. (2021). p. 4.
[107] Laffier, J. & Rehman, A. (2023)., p. 8.
[108] Ibid., p. 5.
[109] Ibid. p. 10.

There are multiple real-life examples I can give on the impact of sexfakes on a victim's life. For example, in 2019, deepfakes portraying extreme acts of sexual violence were distributed on pornographic websites of British poet and novelist Helen Mort.[110] After this incident, she experienced recurring nightmares where these images repeatedly played out and felt intense anxiety about going outside.[111] The threats "literally, albeit not physically, [penetrated her] body".[112]

In another example, Australian high school student Noelle Martin found sexually explicit deepfakes of herself after googling her name out of curiosity.[113] After it became public, she encountered death and rape threats, attempts of sextortion, persistent stalking, and unwelcome sexual advances.[114] Asking for help from authorities was to no avail as there was no legislation in place in Australia that could help her out.[115] The impact these deepfake attacks had on her was long-lasting, affecting her social life, law school prospects, and comfort in public places[116], giving Martin what she calls "a lifelong sentence"[117].

As we can see from the examples, the offences do not stay in the virtual world, they cause violence in the real world as well. Victims are easily identified on the footage and personal information is often posted online together with the material. Consequently, victims are targeted offline as well, being harassed, coerced and extorted by perpetrators.[118]

---

[110] Mort, H. (2022). 'The images are seared onto my retinas - I felt ashamed': The poet who was a victim of deepfake porn. Retrieved from: https://www.telegraph.co.uk/women/life/images-seared-onto-retinas-felt-ashamed-poetvictim-deepfake/ on 8 April 2024;
Laffier, J. & Rehman, A. (2023)., p. 10.
[111] Ibid.
[112] Ibid.
[113] Aitchison, M (2023), 'Aussie student's X-rated horror after innocently Googling her own name to discover someone had done the unthinkable - and her life will never be the same again' retrieved from: https://www.dailymail.co.uk/news/Article-11981501/Aussie-students-horror-Googling-life-never-again.html on 8 April 2024.
[114] Laffier, J. & Rehman, A. (2023)., p. 9-10.
[115] Aitchison, M (2023), 'Aussie student's X-rated horror after innocently Googling her own name to discover someone had done the unthinkable - and her life will never be the same again' retrieved from: https://www.dailymail.co.uk/news/Article-11981501/Aussie-students-horror-Googling-life-never-again.html on 8 April 2024.
[116] Laffier, J. & Rehman, A. (2023)., p. 9-10.
[117] El Atillah, I. (2023), ''Living a lifelong sentence': How AI is trapping women in a deepfake porn hell" retrieved from: https://www.euronews.com/next/2023/04/22/a-lifelong-sentence-the-women-trapped-in-a-deepfake-porn-hell on 8 April 2024.
[118] Laffier, J. & Rehman, A. (2023)., p. 7.

Lastly, separate attention should be given to the forced withdrawal from online spaces by women[119], or as Amnesty International calls it: 'the silencing effect'[120]. When a sexfake is made of a woman, the consequences may be so high for her that as a result, she starts self-censoring her online presence.[121] A report by the National Democratic Institute sheds additional light on the impact silencing and censoring have on women who are subjected to online abuse, noting that:[122]

> *"By silencing and excluding the voices of women and other marginalised groups, online harassment fundamentally challenges both women's political engagement and the integrity of the information space. … In these circumstances, women judge that the costs and danger of participation outweigh the benefits and withdraw from or choose not to enter the political arena at all."*[123]

The self-censoring of women in the online space is highly problematic. Nowadays, a lot of important discussions are held online on a broad range of topics. Representation of different opinions and expertise is therefore necessary, including the need for female representation.[124] This also sends a worrying message to younger generations, implying that women's voices do not matter and are unwelcome in the online sphere (and therefore indirectly also in the offline sphere).[125]

An example of the silencing effect can be seen in the case of Rana Ayyub. Ayyub is an Indian investigative journalist who was the victim of a cruel attack by anonymous netizens[126] who distributed sexfakes of her along with her home address, phone number, and the text "I am available", after her political remarks on the child rape of a Kashmiri girl.[127] Even though she

---

[119] Laffier, J. & Rehman, A. (2023)., p. 8.
[120] Amnesty International (2018), 'Toxic Twitter – the Silencing Effect' retrieved from: https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-Chapter-5-5/#:~:text=It%20State s%2C,integrity%20of%20the%20information%20space%E2%80%A6 on 9 April 2024.
[121] Laffier, J. & Rehman, A. (2023)., p. 10.
[122] Ibid., p. 5.
[123] Gender, Women and Democracy. (2017). #Notthecost: Stopping violence against women in politics. National Democratic Institute, p. 14.; Laffier, J. & Rehman, A. (2023)., p. 5.
[124] Amnesty International (2018), 'Toxic Twitter – the Silencing Effect' retrieved from: https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-Chapter-5-5/#:~:text=It%20State s%2C,integrity%20of%20the%20information%20space%E2%80%A6 on 9 April 2024.
[125] Ibid.
[126] Netizens are active participants in the online community of the Internet (see: https://www.merriam-webster.com/dictionary/netizen).
[127] Maddocks, S. (2020). P. 416-419.

filed a complaint with the Delhi Police as a response to the attacks, nothing was done by the police as they said the culprits could not be identified.[128] The Indian government only undertook action after intervention by the UN, but to this day Ayyub has not mentally recovered from the attacks and has practised self-censorship ever since.[129] She has retreated from multiple online platforms and does not want her picture taken any more out of fear that they will be exploited to generate more sexfakes of her.[130]

On the other hand, there are still women like Kate Isaacs who do not want to give in to the threats. She is the leader of an anti-porn campaign group called 'Not Your Porn', an author and a social and political activist.[131] She conducted research into the activities and regulation of the pornography industry, campaigning for the removal of sexually explicit content hosted on pornographic websites when individuals have not consented to the dissemination of the footage and have not undergone any age verifications.[132] Her activities put a target on her back and she became the victim of deepfake pornography herself.[133] According to Isaacs, the reason for the attack was to 'punish' her for her beliefs, as a retaliation for her 'crime' as an image offence advocate.[134] Fortunately, the abuse didn't stop her from staying in the online space, and she is still active on social media on both her personal[135] and her campaigning[136] Instagram accounts.

2.5. A New Threat Ahead: the Metaverse

2.5.1. The Metaverse Explained

The term Metaverse gained worldwide traction in 2021 when Mark Zuckerberg first announced its new platform 'Meta', but the term itself is not new.[137]

---

[128] Okolie, C. (2023). p. 7; Laffier, J. & Rehman, A. (2023)., p. 9.
[129] Ibid., p. 7.
[130] Ibid.
[131] Ibid.
[132] Ibid.
[133] Ibid.
[134] Ibid.
[135] Personal Instagram Kate Isaacs: https://www.instagram.com/katefisaacs/?hl=nl.
[136] Instagram for the campaign 'not your porn': https://www.instagram.com/notyourpxrn/?hl=nl.
[137] XRToday (2021), 'Unpacking Meta: Where did the Word Metaverse come from?'. Retrieved from https://www.xrtoday.com/virtual-reality/unpacking-meta-where-did-the-word-Metaverse-come-from/ on 18 March.

The term 'Metaverse' was first coined by Neal Stephenson in his book 'Snow Crash' in 1992.[138] In the story, the main character Hiro transitions between the real world and a place called the Metaverse, a digitally coded urban environment where users can engage in lifelike experiences, thereby seeking to escape from a bleak totalitarian reality.[139] This story was not completely based on fiction though. The idea of an immersive digital reality detached from the physical world already existed before 1992.[140] By then, Morton Heilig had created the first VR (Virtual Reality) machine in 1956, and in the 1970s MIT had created the 'Aspen Movie Map', in which users could take a tour around the town of Aspen, Colorado, all while staying seated behind their computer.[141]

Stephenson's invented word the 'Metaverse' comes from the Greek term 'meta' – which means 'after' or 'beyond' – and the English word 'universe'.[142] Characteristics of Stephenson's Metaverse are, among others: "the Metaverse is three-dimensional; the Metaverse is a metaphor for the real-world; users can access the Metaverse using goggles (much like today's VR headsets); users experience the Metaverse from a first-person perspective; the virtual avatars of users are partially customisable".[143]

Since 'Snow Crash', technology has developed rapidly and across the 2000s, games like 'Second Life' and 'The Sims' were created. In these games, a three-dimensional virtual world is generated where users can design a virtual lookalike of themselves, known as avatars, and interact with other avatars, locations or objects.[144] One could argue that these games created a Metaverse before the term became widely recognised, albeit using different terminology.[145]

---

[138] Forbes, 'A Short History of the Metaverse'. Retrieved from: https://www.forbes.com/sites/bernardmarr/2022/03/21/a-short-history-of-the-Metaverse/ on 18 March 2024.
[139] ibid; XRToday (2021), 'Unpacking Meta: Where did the Word Metaverse come from?'. Retrieved from https://www.xrtoday.com/virtual-reality/unpacking-meta-where-did-the-word-Metaverse-come-from/ on 18 March.
[140] Ibid.
[141] Forbes, 'A Short History of the Metaverse'. Retrieved from: https://www.forbes.com/sites/bernardmarr/2022/03/21/a-short-history-of-the-Metaverse/ on 18 March 2024.
[142] XRToday (2021), 'Unpacking Meta: Where did the Word Metaverse come from?'. Retrieved from https://www.xrtoday.com/virtual-reality/unpacking-meta-where-did-the-word-Metaverse-come-from/ on 18 March.
[143] Ibid.
[144] How Stuff Works (2021). How Second Life Works. Retrieved from: https://computer.howstuffworks.com/internet/social-networking/networks/second-life.htm on 23 June 2024.
[145] XRToday (2021), 'Unpacking Meta: Where did the Word Metaverse come from?'. Retrieved from https://www.xrtoday.com/virtual-reality/unpacking-meta-where-did-the-word-Metaverse-come-from/ on 18 March.

In more recent years, books like 'Ready Player One', games like 'Pokémon Go' and 'Fortnite', and home interior apps like those of IKEA, have further merged the virtual and the real world, closely aligning with the original vision of the Metaverse as depicted in 'Snow Crash'.[146]

In 2021, the term 'Metaverse' suddenly gained a lot of new attention, even getting selected for the 'Word of the Year' competition by Collins Dictionary.[147] This recognition was a result of Mark Zuckerberg's announcement that the new parent company of Facebook will be called 'Meta', whose focus will be to bring the Metaverse to life.[148] Meta's Metaverse will not be a completely separate online world but will "feel like a hybrid of today's online social experiences, sometimes expanded into three dimensions or projected into the physical world. It will let you share immersive experiences with other people even when you can't be together — and do things together you couldn't do in the physical world."[149] Zuckerberg views the Metaverse as something going beyond current 2D social interactions, just as the Greek etymology of the word implies.[150]

In general, Ng distinguishes four key components of the Metaverse: immersion, advanced computing, socialisation and decentralisation (see Figure 3).[151]

Firstly, immersive technology blurs the lines between the offline and online world, enabling users to fully engage through augmented or virtual reality (See Figure 2). The Metaverse with augmented reality would be partially virtual whereas the use of virtual reality could create a fully virtual world.[152]

---

[146] Forbes, 'A Short History of the Metaverse'. Retrieved from: https://www.forbes.com/sites/bernardmarr/2022/03/21/a-short-history-of-the-Metaverse/ on 18 March 2024.
[147] XRToday (2021), 'Unpacking Meta: Where did the Word Metaverse come from?'. Retrieved from https://www.xrtoday.com/virtual-reality/unpacking-meta-where-did-the-word-Metaverse-come-from/ on 18 March.
[148] Meta (2021), Introducing Meta: A Social Technology Company. Retrieved from: https://about.fb.com/news/2021/10/facebook-company-is-now-meta/ on 23 June 2024.
[149] Meta (2021), 'Introducing Meta: A Social Technology Company'. Retrieved from: https://about.fb.com/news/2021/10/facebook-company-is-now-meta/ on 23 June 2024.
[150] XRToday (2021), 'Unpacking Meta: Where did the Word Metaverse come from?'. Retrieved from https://www.xrtoday.com/virtual-reality/unpacking-meta-where-did-the-word-Metaverse-come-from/ on 18 March.
[151] Ng, D. T. K. (2022). p. 198.
[152] Ibid.

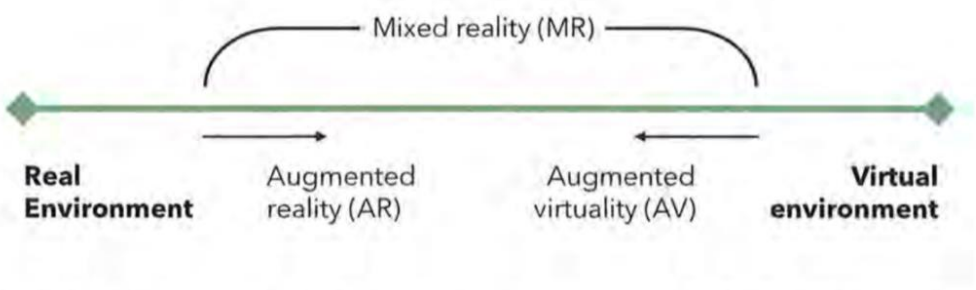


Figure 2: Mr. Dr. Bart W. Schermer & Joas van Ham MSc (2021). 'Regulering van immersieve technologieën Wetenschappelijk Onderzoek- en Documentatiecentrum' retrieved from: https://open.overheid.nl/documenten/ronl-d81ef4594c8bcca4b8e04ec659d0b6930f2cad67/pdf

Secondly, the Metaverse can only be created through advanced computing technologies. AI, data mining, new servers with high bandwidth, etc. are all needed to create an immersive world. This technology was not as advanced yet previously but we are more and more overcoming the technological challenges involved with the Metaverse.[153]

Thirdly, the Metaverse gives users the possibility to socialise with each other.[154] The metaverse can be used in diverse contexts to enable users to create their own avatars who then can share their thoughts and ideas with other avatars, providing a highly immersive virtual experience.[155]

Lastly, the Metaverse is decentralised. There is no centralised authority managing the whole Metaverse.[156] Tim Sweeney, CEO of Epic Games, is an advocate for an open Metaverse and has stated that the Metaverse cannot be constructed or owned by one single company, just as how it is not possible to have one company or government owning the internet.[157] Decentralisation would also guarantee that the Metaverse is ongoing and ever-present, meaning it will continue to evolve even when users are not logged in.[158]

[153] Ng, D. T. K. (2022). p. 198.
[154] Ibid, p. 194.
[155] Kye, B., et. al. (2021). p. 10.
[156] Ng, D. T. K. (2022). p. 198.
[157] Dolan, L. (2022). p. 13.
[158] Ibid.

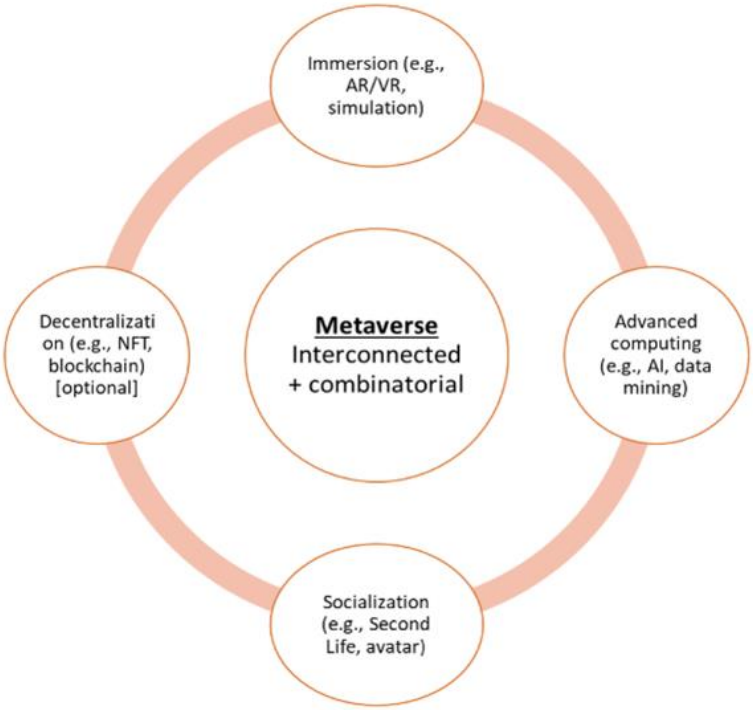


Figure 3: Ng, D. T. K. (2022). p. 200.

Important to note is that it is always *the* Metaverse and not *a* Metaverse.[159] This concept, as envisioned by figures like Neal Stephenson, Mark Zuckerberg, and others, describes a universal, cohesive, and interoperable 3D space that will integrate the numerous virtual worlds that exist today.[160] Many companies are currently developing a version or feature of the Metaverse of which Roblox, Microsoft Mesh, and Nvidia Omniverse are just some examples.[161]

Another side note to make is that, even though the possibilities for the Metaverse are endless, currently no 'real' Metaverse exists.[162] However, experts believe that the Metaverse will emerge as a viable technology product in the near future.[163]

---

[159] XRToday (2021), 'Unpacking Meta: Where did the Word Metaverse come from?'. Retrieved from https://www.xrtoday.com/virtual-reality/unpacking-meta-where-did-the-word-Metaverse-come-from/ on 18 March.
[160] Ibid.
[161] Cubewealth (2023), 'what is the Metaverse: Origins, Platforms, Future, Warnings?' retrieved from: https://www.bankoncube.com/post/what-is-the-Metaverse on 18 March 2024.; Tariq, S. et al. (2023). p. 16.
[162] Europol (2022)., p. 7-8.; Cubewealth (2023), 'what is the Metaverse: Origins, Platforms, Future, Warnings?' retrieved from: https://www.bankoncube.com/post/what-is-the-Metaverse on 18 March 2024.
[163] XRToday, 'Unpacking Meta: Where did the Word Metaverse come from?'. Retrieved from https://www.xrtoday.com/virtual-reality/unpacking-meta-where-did-the-word-Metaverse-come-from/ on 18 March.

2.5.2. Sexfakes in the Metaverse

Even though the Metaverse as envisioned does not exist yet, significant risks are attached to immersing oneself in the Metaverse. One of them is the risk of impersonation.[164] A central point to the Metaverse is that users can create an avatar of themselves that copies the user's mindset, behaviour, voice, and sometimes appearance (the latter is not always necessary).[165] As Tariq describes it: "The Metaverse will be so fake that fake will be real".[166] However, in the Metaverse, chances are that certain users do not want to use their own physical appearance and identity but instead want to create an avatar based on a completely different person (with or without permission).[167] The latter is in fact easier than in the offline world, as the abundance of personal information available on the internet makes the creation of a realistic digital clone of another user a piece of cake for a bit of a tech-savvy user.[168] Once a user has made a digital twin of someone else, deepfake material of the avatar is easily made, including sexfakes.[169]

Women should therefore be concerned about their personal safety in the Metaverse.[170] Even though the 'real' Metaverse does not exist yet, considering the ambitions for its development, the goal of transitioning our daily activities into it, and the uncertainty regarding how it will alter our lives, I believe it is quintessential to prepare for what the future may hold.[171] An extensive analysis of how sexfakes in the Metaverse should be regulated is therefore needed.

2.6. Conclusion

In this chapter, I discussed the definition of deepfakes, its benefits and risks, with a particular focus on sexfakes and their disproportionate impact on women and girls. This gendered

[164] Tariq, S. et al. (2023). p. 16.
[165] Ibid; Levy, S. (2022) 'What's Deepfake Bruce Willis Doing in My Metaverse?' retrieved from: https://www.wired.com/story/plaintext-bruce-willis-deepfake-Metaverse/ on 18 March 2024.
[166] Ibid.
[167] Tariq, S. et al. (2023). p. 16.
[168] Tariq, S. et al. (2023). p. 17-18.
[169] Ibid., p. 16.
[170] Marr, B. (2024). 'The Metaverse And Its Dark Side: Confronting The Reality Of Virtual Rape' retrieved from: https://www.forbes.com/sites/bernardmarr/2024/01/16/the-Metaverse-and-its-dark-side-confronting-the-reality-of-virtual-rape/ on 18 March 2024.
[171] Rigotti, C., & Malgieri, G. (2023). p. 14.

dimension is not incidental but reflects and reinforces existing power dynamics and the objectification of women in both online and offline spaces.

While sexfakes initially targeted female celebrities, the rapid evolution and accessibility of AI technology now enables anyone to create and distribute a sexfake of their classmate/colleague/ex-partner/ etc. without the depicted person's consent. Beyond the immediate psychological trauma and reputational damage, sexfakes contribute to a broader 'silencing effect,' pushing women out of online spaces and limiting their participation in digital discourse. This withdrawal has far-reaching implications for gender equality in the digital age. The stories of victims like Helen Mort, Noelle Martin and Rana Ayyub underscore the impact that sexfakes have not only in the digital sphere but also in the real world.

Looking ahead, as the Metaverse promises to revolutionise digital interaction in a 3D capacity, it will also introduce new dimensions of vulnerability for women. The ability to create immersive, lifelike digital avatars opens the door to unprecedented levels of identity manipulation and sexploitation. It is, therefore, necessary to intervene now, before it is too late. However, focusing solely on their novelty risks clouding deeply engrained social norms, ideologies, and ongoing struggles of women in the offline world.[172] The misogyny leaking through from the real world into the virtual world should be taken into account when looking for solutions. Any effective regulation must tackle both the symptoms and the root causes of sexfakes.

In the next chapter, I shall analyse which European regulatory instruments are already in place for the protection of women and girls from sexfakes, evaluating their effectiveness in addressing this gendered harm and identifying areas where further action is needed.

[172] Rigotti, C., & Malgieri, G. (2023). p. 14.

Chapter 3: Regulatory Framework on Deepfakes

3.1. Introduction

In this chapter, I shall delve into the regulatory framework of both the EU and the CoE. Section 3.2. focuses on the CoE instruments and section 3.3. on the relevant EU legislation. Scholars have often argued that deepfakes are a violation of the depicted person's identity, making it a privacy issue.[173] Even though this is true, I shall also argue that the focus on privacy alone does not suffice.[174] In my opinion, there should be a shift of focus from privacy protection to a focus on the gendered harm of sexfakes. Therefore, I will also extensively analyse the EU's and CoE's framework on gender-based violence.

3.2. Regulatory Framework of the Council of Europe on Sexfakes

The CoE has multiple instruments that could be applicable in the case of non-consensual sexfakes. I shall first elaborate on the CoE Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, and the Convention on Cybercrime, after which I will discuss the relevant provisions of the ECHR. Afterwards, I shall specifically delve into the CoE's gender-based violence framework. The Istanbul Convention will be discussed together with GREVIO Recommendation No. 1 as the latter interprets the IC in the context of cyber violence.

3.2.1. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law

The CoE Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law was adopted on 17 May 2024 by the Committee of Ministers of the Council of Europe and will be opened for signature on 5 September 2024, at the Conference of Ministers of Justice in Vilnius.[175]

---

[173] McGlynn, C., et. al. (2017). p. 36.
[174] Ibid.
[175] CoE. (2024). Committee on Artificial Intelligence (CAI). Retrieved from: https://www.coe.int/en/web/artificial-intelligence/cai on 18 June 2024.

The Convention addresses actions within the lifespan of AI systems that could potentially interfere with human rights, democracy and the rule of law.[176] Consequently, Parties will have the obligation to take measures to ensure that the actions related to the AI systems' life cycle comply with the obligations that have been set forth by international human rights law and national legislation.[177] If one of these rights has been violated, Parties have to offer remedies.[178] The Convention provides for a risk and impact management framework that prescribes to Parties that they should adopt or maintain measures for the identification, assessment, prevention and mitigation of risks posed by AI systems.[179] Parties should consider both actual and potential impacts on human rights, democracy, and the rule of law.[180] This assessment should take into account the context and intended use of AI systems, as well as the severity and probability of potential impacts.[181] Where appropriate, parties should also consider the perspectives of relevant stakeholders, particularly those whose rights may be affected.[182] If the assessment reveals that the AI system poses significant risks, a party can impose a moratorium or ban the AI system in question.[183]

Even though the new Framework Convention on Artificial Intelligence seems to be the most straightforward convention to look at when it comes to deepfakes, it actually does not provide much help in finding a solution for the protection of victims against sexfakes. The Convention's wording is very broad and lays a lot of responsibility with the Parties to, firstly, assess the risk, and secondly, decide on what to do if an AI system poses a significant risk to human rights. In my opinion, I do not think that this new Framework Convention will add anything new to the table as it reiterates that the AI systems have to adhere to the already existing national and international frameworks. But what these obligations are, is not made clear. In the specific case of sexfakes, the Convention does not mention the words 'deep fake' or 'image-based sexual violence' or 'sexual abuse' even once in its preambles or Articles. The explanatory

[176] Article 3 Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.
[177] Article 4 and 14 Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.
[178] Ibid.
[179] Article 16 Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.
[180] Ibid.
[181] Ibid.
[182] Ibid.
[183] Article 16(4) Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.

memorandum to the Convention does not refer to deepfakes either and the only time sexual abuse is mentioned is in the context of children.[184]

Furthermore, the Convention will only be opened for signature on 5 September 2024.[185] As this thesis will be finished and handed in before September 2024, it has to be seen how many Parties will eventually sign and ratify the Convention as well as the long-term impact of the Convention.

3.2.2. Convention on Cybercrime

The Convention on Cybercrime of the Council of Europe (the Budapest Convention) is a legally binding treaty focused on addressing cybercrime and electronic evidence.[186] It entered into force on 1 July 2004 and currently has 75 Parties.[187] Along with the Convention, the Cybercrime Convention Committee (T-CY) was founded.[188] The T-CY monitors "the effective use and implementation of the Convention, the exchange of information and consideration of any future amendments".[189]

The Budapest Convention is the only international treaty that criminalises behaviours and activities carried out through computer systems and information networks.[190] Parties to the Convention are obliged to criminalise offences involving computer data and systems, encompassing acts such as the production, distribution or possession of child sexual abuse material, and infringements of copyright and related rights.[191] Furthermore, when a cybercrime

---

[184] The Committee on Artificial Intelligence states that: "*In view of the serious risk that artificial intelligence technologies could be used to facilitate sexual exploitation and sexual abuse of children, and the specific risks that it poses to children, in the context of implementation of this provision, the Drafters considered the obligations set forth in the Lanzarote Convention, the Optional Protocol to the UN Convention on the Rights of the Child on the sale of children, child prostitution and child pornography, and General Comment No. 25 to the UNCRC on children's rights in relation to the digital environment.*" (Committee on Artificial Intelligence (2024), 'Draft Framework Convention on artificial intelligence, human rights, democracy and the rule of law: Draft Explanatory Report' CM(2024)52-addprov, par 118).

[185] CoE. (2024). Committee on Artificial Intelligence (CAI). Retrieved from: https://www.coe.int/en/web/artificial-intelligence/cai on 18 June 2024.

[186] Van der Wilk, A. (2021)., p. 7.

[187] CoE, 'The Budapest Convention (ETS No. 185) and its Protocols'. Retrieved from: https://www.coe.int/en/web/cybercrime/the-budapest-convention on 23 June 2024.

[188] Article 46 Budapest Convention.

[189] CoE, 'Cybercrime Convention Committee'. Retrieved from: https://www.coe.int/en/web/cybercrime/tcy on 23 June 2024.

[190] Velasco, C. (202). p. 116-117.

[191] Van der Wilk, A. (2021). p. 7.

has been reported, the Parties must have established efficient procedures and powers to make sure that electronic evidence is gathered and saved for a criminal investigation.[192] They should also facilitate international cooperation and mutual assistance in these proceedings.[193]

However, just like the Framework Convention, the Budapest Convention does not effectively protect sexfake victims either. The original drafters anticipated any future changes in the cybercrime landscape by drafting the Convention with 'technological neutrality' in mind. This would mean that the Articles of the Budapest Convention would still apply in the case of deepfakes, even when they are not explicitly mentioned. However, the T-CY has not taken the opportunity yet to clarify whether the misuse of deepfakes, and more specifically sexfakes, could fall under the Budapest Convention. So far, it has acknowledged that threatening behaviour can occur in the virtual world through chat rooms, social networking sites, etc.[194] In its Mapping study on cyber violence, the T-CY also considered that cyberbullying is an umbrella term for many bullying activities, including the non-consensual creation and distribution of intimate images or videos, and sextortion.[195] The T-CY reflected on how data on cyber violence against women is lacking.[196] Even though these observations have been considerate of the image-based sexual abuse and the gendered nature of cyberbullying, we cannot say for certain whether fake content like sexfakes would fall under the same umbrella term. And even if the T-CY would clarify that sexfakes fall under the term cyberbullying, I would argue that this does not suffice. The grave impact sexfakes have on the lives of victims cannot be considered as mere 'bullying'. Instead, it is a form of sexual violence which should be taken very seriously.

3.2.3. European Convention on Human Rights

The European Convention on Human Rights (ECHR) stands as the foundational treaty of the Council of Europe and forms the basis for all its activities.[197] Adopted in 1950 and entering into

[192] Van der Wilk, A. (2021). p. 7.
[193] Ibid.
[194] T-CY (2018). p. 24.
[195] Ibid., p. 7 & 11.
[196] Ibid., p. 16.
[197] CoE. 'A Convention to protect your rights and liberties.' Retrieved from: https://www.coe.int/en/web/human-rights-convention#:~:text=The%20European%20Convention%20on%20Human%20Rights%20is%20the%20first%20Council,entered%20into%20force%20in%201953 on 12 July 2024.

force in 1953, its ratification is mandatory for any State seeking membership of the CoE.[198] The European Court of Human Rights (ECtHR), located in Strasbourg, is responsible for ensuring that the 46 MS of the CoE adhere to the Convention's principles.[199] In the following part I shall consecutively discuss the ECHR Articles that I consider relevant to the topic of sexfakes. First, Article 8, right to respect for private and family life, will be discussed, after which Article 10, freedom of expression, and Article 3, right to be free from degrading, will be examined.

Article 8: Right to respect for private and family life

Article 8 ECHR states that everyone has the right to respect for his private and family life, his home and his correspondence. This entails both the right to someone's physical integrity and the right to privacy. I will first discuss the right to physical integrity and then the right to privacy.

The ECtHR first acknowledged that the notion of private life also covered the physical and moral integrity of the person in the case of *X and Y v. the Netherlands*.[200] This, as "a person's body concerns the most intimate aspect of private life".[201] In the case of cyber violence, Article 8 is applicable because such violence endangers bodily integrity and the right to a private life.[202] Parties have a duty to safeguard an individual's physical and moral integrity from violations by others.[203]

I would argue that the creation and dissemination of non-consensual sexfakes is a grave violation of someone's personal integrity. It doesn't matter here that the footage depicts an act which technically has not been performed by the victim. The fact that other people think she is engaging in those acts with her own body would in my opinion suffice to conclude that the victim's personal integrity has been violated.

[198] CoE. 'A Convention to protect your rights and liberties.' Retrieved from: https://www.coe.int/en/web/human-rights-convention#:~:text=The%20European%20Convention%20on%20Human%20Rights%20is%20the%20first%20Council,entered%20into%20force%20in%201953 on 12 July 2024.
[199] Ibid.
[200] ECtHR *X and Y v. the Netherlands*, para. 22.
[201] Ibid; see also: ECtHR *Y.F. v. Turkey*, para. 33.
[202] ECtHR *Milićević v. Montenegro*, paras. 54-56; ECtHR *E.S. and Others v. Slovakia*, para. 44.
[203] ECtHR *Buturugă v. Romania*, paras. 74, 78-79; ECtHR *Volodina v. Russia (no. 2)*, paras. 48-49.

The right to privacy can be divided into multiple 'sub rights'. In this section, I will consecutively discuss the right to one's image, the right to protection of one's reputation, and the right to be forgotten (RTBF).

The ECtHR has ruled that: "'private life' is a broad term not susceptible to exhaustive definition, which covers the physical and psychological integrity of a person and can therefore embrace multiple aspects of a person's identity, such as . . . elements relating to a person's right to their image[204]".[205] In *Von Hannover v. Germany* the ECtHR asserted that the right to the protection of one's image is "one of the essential components of personal development and presupposes the right to control the use of that image".[206] The term 'image' is interpreted broadly, encompassing not only portraits, photographs or videos depicting a person, but also their 'likeness' or resemblance.[207] The mere recognition of a person could be enough to invoke someone's image rights.[208] Deepfakes exploit various different angles and aspects of a person's identity and image, manipulating not only one's facial features, but also mannerisms, and speech patterns.[209] Even though the Court has not taken a stance on deepfakes yet, its broad reading of Article 8 ECHR could indicate that victims of sexfakes could rely on the Court to conclude that manipulated videos harm their image.[210]

Article 8 ECHR also protects the right to protection of reputation. The Court in *Axel Springer AG v. Germany* stated that in order to have a violation of Article 8, an attack on an individual's reputation has to "attain a certain level of seriousness and in a manner causing prejudice to personal enjoyment of the right to respect for private life".[211] This criterion applies to both professional and social reputation.[212] Furthermore, a clear link between the applicant and the purported attack on their reputation has to be proven.[213] The Court considers how well-known an applicant was when the alleged defamatory comments were made, noting that public figures are subject to a wider range of accepted criticism compared to ordinary citizens, and also considers the content of the statements made.[214]

---

[204] ECtHR *Axel Springer AG v. Germany*, para. 83.
[205] Boyd, P. (2022). p. 527-528.
[206] ECtHR *Von Hannover v. Germany (no. 2)*, para 96.
[207] Huijstee, M.V. et al. (2021), p. 40.
[208] Ibid.
[209] Boyd, P. (2022). p. 525.
[210] Ibid.
[211] ECtHR *Axel Springer AG v. Germany*, paras. 83-84.
[212] ECtHR *Denisov v. Ukraine*, para. 112.
[213] ECtHR *Putistin v. Ukraine*, para. 40.
[214] ECtHR *Jishkariani v. Georgia.*

Once again, the Court has not pronounced its stance on non-consensual sexfakes yet but it has emphasised that in the context of the internet, the level of seriousness test is important.[215] For example, *Tamiz v. the United Kingdom* and *Çakmak v. Turkey* concerned the posting of hurtful comments online.[216] Even though the Court acknowledged that these comments could be considered as offensive or even defamatory, it concluded that most of the comments were probably too minor in nature, and/or their dissemination too limited, for them to inflict any substantial harm to someone's reputation.[217] In the case of sexfakes, however, I would argue that the 'level or seriousness' test is met. As explained in Chapter 2, the impact that sexual deepfakes have on victims is detrimental as they tarnish a woman's reputation, forcing her to switch jobs sometimes and even losing friendships over it.[218] In the case of Rana Ayyub, the sexfakes were used to discredit her reputation of being a good investigative journalist. The dissemination of sexfakes cannot be seen as mere offensive comments on the internet as was the case in *Tamiz v. the United Kingdom* and *Çakmak v. Turkey,* as it is way more than that. Therefore, I would conclude that the creation and dissemination of non-consensual sexfakes should result in a violation of Article 8 ECHR, the right to protection of reputation.

Lastly, the right to be forgotten. The Court has ordered multiple times that the identity of offenders should be anonymised to respect their RTBF after a certain period of time has elapsed.[219] In my opinion, not only offenders but also victims should be able to rely on the RTBF under Article 8 ECHR when non-consensual sexfakes are being shared on online platforms. However, as we will see in section 3.3., EU law, Article 17 GDPR will provide a more helpful tool to have one's information erased. I will therefore discuss the right to be forgotten more extensively in the next section under EU law.

Article 10: Freedom of Expression

Article 10 of the Convention protects everyone's right to freedom of expression. It is one of the core rights of the ECHR and it protects even 'information' or 'ideas' that "offend, shock or

[215] ECtHR *Tamiz v. the United Kingdom*, paras. 80-81; *Çakmak v. Turkey*, paras. 42, 50.
[216] Ibid.
[217] ECtHR (2016), p. 52.
[218] Laffier, J. & Rehman, A. (2023)., p. 5-8.
[219] ECtHR *M.L. and W.W. v. Germany*, para. 100; ECtHR *M.L. v. Slovakia*, para. 38.

disturb the State or any sector of the population"[220].[221] Article 10 does not restrict the forms and means through which information and ideas can be generated, shared, and received.[222] Consequently, all methods of expression, including those involving deepfakes, are protected under Article 10 of the Convention.[223]

A party to the Convention has both a positive and a negative obligation to protect the freedom of expression.[224] Under its positive obligation, the State must create an environment that enables the enjoyment of the freedom of expression.[225] Per its negative obligation, a State must refrain from interfering with the exercise of this freedom within its prescribed boundaries.[226]

The State has a positive obligation in my opinion to fight against the 'silencing effect' caused by non-consensual sexfakes. As mentioned in Chapter 2, non-consensual sexfakes have the consequence that women are being 'silenced' in the online sphere as they self-censor their presence online and withdraw from the digital world. This is a very disturbing trend which has to be challenged by the States to the best they can.

On the other hand, the State also has a negative obligation not to interfere with the freedom of expression (of for example the deepfake creators). It can only restrict this right in certain situations. Paragraph 2 of Article 10 prescribes that the right to freedom of expression can only be restricted when: there is a legal basis, a legitimate aim, and when this is necessary in a democratic society. The latter consists of a twofold test: (1) there has to be a reasonable relationship of proportionality between, on the one hand, the restrictions imposed by national law on the applicant and, on the other hand, the legitimate aim pursued, and (2) there were no less intrusive measures available for achieving the same goal. In *Observer and Guardian v. the United Kingdom*, the ECtHR interpreted the term 'necessary' in paragraph 2 of Article 10 as requiring a 'pressing social need'.[227] Whether or not a pressing social need existed for a State to intervene will depend on the circumstances of the specific case.

In my opinion, States should be able to restrict a user's freedom to create deepfakes in cases where those deepfakes are sexually explicit and non-consensual. The legal basis would be the domestic implementation of Article 8 ECHR and the legitimate aim the protection of reputation

[220] ECtHR *Handyside v. the United Kingdom*, para. 49.
[221] CoE, 'Hate Crime and Hate Speech'. Retrieved from: https://rm.coe.int/thematic-factsheet-hate-crime-eng-docx/1680a96865 on 8 July 2024.
[222] McGoldrick, D. (2013). p. 126.
[223] Mammadzada, I. (2021). p. 24.
[224] Ibid., p. 25.
[225] Ibid.
[226] Ibid.
[227] ECtHR *Observer and Guardian v. the United Kingdom.*

and rights of others. The restriction would also be necessary in a democratic society as the banning of non-consensual sexfakes is necessary to protect the reputation and rights of others. No less intrusive measures are possible either. The most intrusive measure would be to ban all types of deepfakes, regardless of their content. If only non-consensual sexfakes are banned, which objectively go beyond the premise of the protection of freedom of expression, the State has used its least intrusive measure at hand.

Article 3: Degrading Treatment

Lastly, I would argue that sexfake perpetrators violate the victim's right to be free from degrading treatment, Article 3 ECHR. Treatment is deemed 'degrading' if it humiliates or debases an individual, disregards or diminishes their human dignity, or induces feelings of fear, anguish or inferiority that can undermine their moral and physical resilience.[228] It is enough for the victim to feel humiliated in their own eyes, regardless of the perception of others.[229] Additionally, while the intent to humiliate or degrade the victim is a relevant factor, the absence of such intent does not automatically negate a violation of Article 3 ECHR.[230] Considering the grave impact sexfakes have on a woman's life, I would argue that the creation and dissemination of this content would fall under Article 3 ECHR.

However, the chances are small that the ECtHR would consider non-consensual sexually explicit deepfakes under Article 3. This can be derived from the case of *Buturugā v. Romania.* Mrs Buturuga reported to the police that, besides being the victim of domestic violence at the hands of her former husband, he had also accessed her Facebook account and copied her private conversations, documents and photos.[231] She asked the police to undertake an electronic search of their family computer in order to obtain evidence of the offence.[232] However, the police dismissed this particular part of her complaint, stating that it was unrelated to the allegations of domestic violence.[233] Consequently, Mrs. Buturuga filed another complaint, claiming that her ex-husband had violated the confidentiality of her correspondence. Despite her efforts, the public prosecutor's office dropped the case, deeming her ex-husband's actions insufficiently

[228] ECtHR *Gäfgen v. Germany*, para. 89; ECtHR *Ilaşcu and Others v. Moldova and Russia*, para. 425; ECtHR *M.S.S. v. Belgium and Greece*, para. 220.
[229] Ibid.
[230] Ibid.
[231] CoE. 'International case law'. Retrieved from: https://www.coe.int/en/web/cyber violence/international-case-law#{%22212947575%22:[0]} on 1 July 2024.
[232] Ibid.
[233] Ibid.

serious to be criminal and imposed an administrative fine instead.[234] When Mrs. Buturuga contested this decision in court, the court upheld the dismissal, arguing that the breach of her correspondence was irrelevant because the information was publicly accessible on social media. Eventually, Mrs. Buturuga filed a complaint with the ECtHR.[235] The ECtHR concluded that there was a violation of both Articles 3 and 8 ECHR. However, the ECtHR came to this conclusion by splitting its reasoning into two separate parts.[236] In the first part, the Court examined whether the physical abuses by Mrs Buturuga's ex-husband could constitute inhuman treatment under Article 3.[237] In the second part, the Court assessed whether the violation of the secrecy of correspondence has led to a breach of Article 8 ECHR, right to privacy (so not under Article 3 ECHR).[238] This examination eventually led to the conclusion that Article 8 was violated. The Court's analysis in Buturuga suggests that cyber violence is a matter to be addressed under Article 8 of the ECHR, which is distinct from physical violence that falls under Article 3.[239] If a case of non-consensual sexfakes would come before the Court, it is to be seen what route it will take in finding a violation.

One thing about the Buturuga case has to be noted though. In the next section, I shall delve into the Istanbul Convention, which is specifically designed to tackle gender-based violence. However, when it comes to the recognition of cyberbullying as a form of gender-based violence, the case of Buturuga is revolutionary[240]. The ECtHR explicitly confirms that cyber harassment, cyber violence and malicious impersonation all constitute gender-based violence as they can undermine a woman's psychological and psychical integrity due to her vulnerable position in society.[241] Parties to the Convention have a positive obligation to establish and effectively enforce a legal framework that penalises all forms of domestic violence, including online violence, and to ensure comprehensive protection of victims.[242] The Court's judgment represents a significant advancement in the CoE framework in recognising cyber violence as a continuation of gender-based violence, as well as a human rights violation, having an effect which goes beyond just this Romanian case.[243]

---

[234] CoE. 'International case law'. Retrieved from: https://www.coe.int/en/web/cyber violence/international-case-law#{%22212947575%22:[0]} on 1 July 2024.
[235] Ibid.
[236] Zotti, S. (2023). p. 86-87.
[237] Ibid.
[238] Ibid.
[239] Ibid.
[240] See also: ECtHR *Volodina v. Russia (No.2).*
[241] Zotti, S. (2023). p. 84-85.
[242] Ibid;
[243] Ibid.

3.2.4. Istanbul Convention & GREVIO General Recommendation No. 1

The Istanbul Convention has been coined a landmark treaty for women's rights as it encompasses an extensive set of measures for Parties to prevent and combat all forms of violence against women and domestic violence, positioning violence against women as a human rights violation.[244] The achievement of gender equality is linked with the erasure of this type of discrimination.[245] The Convention is organised around the '4 Ps': "prevention, protection and support of victims, prosecution of offenders and co-ordinated policies".[246] The IC is monitored by the Group of Experts on Action against Violence against Women and Domestic Violence (GREVIO.[247] GREVIO prepares and releases reports assessing the actions taken by State Parties to fulfil the obligations under the Convention.[248] Sometimes it is necessary for GREVIO to step in when a serious, widespread or persistent pattern of violence is happening in a Party.[249] In that case, GREVIO can initiate a special inquiry procedure. Besides the inquiry procedure, GREVIO is also qualified to adopt general recommendations on themes and concepts which are part of the Convention.[250]

Article 3 IC, defines violence against women as covering "all acts of gender-based violence that result in, or are likely to result in, physical, sexual, psychological or economic harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life".[251] Under the Convention, gender-based violence 'shall mean violence that is directed against a woman because she is a woman or that affects women disproportionately'.[252] Article 4(3) confirms that the implementation of the Convention must be secured without discrimination on any ground.[253]

[244] Van der Wilk, A. (2021)., p. 13.
[245] Ibid.
[246] Ibid.
[247] Article 66 Istanbul Convention.
[248] CoE. 'GREVIO'. Retrieved from: https://www.coe.int/en/web/istanbul-convention/grevio on 24 June 2024.
[249] Ibid.
[250] Ibid.
[251] Article 3(a) Istanbul Convention.
[252] Article 3(d) Istanbul Convention.
[253] Article 4(3) Istanbul Convention.

These definitions are especially relevant in cases of cyber violence against women as they cover both psychological and economic harm and not just physical ones.[254]

Despite the relatively young age of the Convention – it entered into force in 2014 – the Convention does not mention cyber, online or digital forms of violence against women, nor does the Explanatory Report to the Convention.[255] The latter only briefly mentions 'online stalking'.[256] This is especially striking as the Budapest Convention was adopted ten years before the Istanbul Convention.[257] How is it possible that the Istanbul Convention did not take into account gender-based violence in the digital sphere? According to the CoE, cyber violence has always been intended to be included in the Convention from the start.[258] To clarify matters for the State Parties, however, GREVIO adopted its first General Recommendation on the Digital Dimension of Violence Against Women in 2021. The Recommendation stresses that there is an overlap between online and offline abuse.[259] Cyber violence against women from offline violence cannot be separated as the harm is not experienced in a vacuum.[260] Harmful behaviour in the digital world disproportionately targets women and girls and is a central element of their experiences of gender-based violence.[261] GREVIO iterates that in its understanding of the concept of the digital dimension of violence against women, both online and technology-facilitated aspects are included. Online aspects include "activities performed and data available on the internet, including internet intermediaries on the surface web as well as the dark web" and technology-facilitated aspects include "activities carried out with the use of technology and communication equipment, including hardware and software".[262]

GREVIO also clarifies its position on non-consensual sexfakes.[263] The non-consensual taking, producing or procuring of intimate images or videos is considered to fall under Article 40 of the Convention, sexual harassment. According to GREVIO, this entails among others, "producing digitally altered imagery in which a person's face or body is superimposed or

[254] EDVAW (2022). p. 15-16.
[255] Guney, G. (2022), 'The Istanbul Convention: A Missed Opportunity in Mainstreaming Cyber violence against Women in Human Rights Law?' retrieved from: https://www.ejiltalk.org/the-istanbul-convention-a-missed-opportunity-in-mainstreaming-cyber violence-against-women-in-human-rights-law/ on 2 May 2024.
[256] Ibid.
[257] Ibid.
[258] EDVAW (2022). p. 15-16.
[259] Ibid.
[260] Ibid.
[261] Ibid.
[262] GREVIO (2021), para. 38.
[263] EDVAW (2022). p. 15-16;

"stitched into" a pornographic photo or video, known as "fake pornography" (such as "deepfakes", when synthetic images are created using artificial intelligence)".[264] The complementary offence of online sextortion also falls under Article 40 of the Convention.[265]

GREVIO's Recommendation also specifically mentions the silencing effect of women due to online abuse. It states:

> *"13. In addition, it has severe implications for women's participatory rights online. The hateful abuse to which women are subjected in online environments causes many women to withdraw from participating online, including from expressing their views on online platforms. This is particularly problematic for women and girls' human rights' defenders, journalists or those in politics, but also for social media influencers or others active on social media and/or in public. Violence against women and girls in the digital sphere thus silences their voices and reduces their perspectives in public debate. As such, GREVIO considers it to not only amount to gender-based violence against women but to undermine a number of other human rights of women as protected by international law."*[266]

GREVIO's recommendation has therefore tackled a lot of issues concerning online gender-based violence and it specifically clarified that non-consensual sexfakes fall under the protection of the IC. However, I still want to highlight some issues that come with the implementation of the IC by State Parties.

Implementation is necessary to ensure that the commitments made at the international level are truly realised in practice within the State. National laws and policies should be aligned with the Convention's provisions. Two issues can be recognised in the context of the IC.

The first one concerns, ironically, the gender-neutral wording of Article 40 IC. The IC employs gender-neutral language in most of the substantive criminal offences, except for the ones on female genital mutilation (FGM), forced sterilisation and forced abortion.[267] These are

[264] GREVIO (2021), para. 38.
[265] Ibid.
[266] Ibid.
[267] Leye, E., et. al. (2021). p. 5.

explicitly framed as primarily affecting women, specifically their sexual and reproductive integrity.[268] Even though the criminal law provisions were drafted in a gender-neutral manner, it was also made clear that this should not prevent the introduction of gender-specific provisions by State Parties.[269] In practice, however, multiple State Parties implemented the substantive criminal offences through only gender-neutral policies and legislation.[270] While criminalising violent acts regardless of the victim's gender is crucial, there's an argument to be made for acknowledging the gendered nature of certain offences (like the disproportionate effect of sexfakes on female victims) and making it an aggravating factor in criminal cases.[271] Otherwise you will get undesirable situations like in Norway, where the country's gender-blind approach to shelters for domestic violence victims resulted in an equal allocation of shelters for men and women, leading to underutilised men's facilities and inefficient resource distribution (22 of the 51 shelters were designated for men, of which 10 remained unused due to insufficient demand).[272] This case illustrates how gender-neutral policies can inadvertently impact the practical handling of gender-based violence and domestic abuse cases.[273]

No surprise then, that GREVIO's first general report condemned this approach.[274] According to GREVIO, a gender-neutral approach "fails to address the specific experiences of women that differ significantly from those of men thus hindering their effective protection".[275] The report argues that gender-neutral policies overlook the power imbalances between genders, obscure accountability for abuse, and impede effective criminal prosecution. Furthermore, this approach can negatively affect service provision, as evidenced by the shortage of women's shelters in Norway.[276]

The second issue concerns the unsatisfactory implementation by State Parties of the offence of sexual harassment under Article 40 IC. While some States have some form of legal prohibition against sexual harassment in place, it is often limited to the context of the workplace.[277] Outside of the workplace, sexual harassment is tackled in a fragmented way, falling under the offences of stalking or indecent assault, indecency, harassment, sexual intimidation or other similar

[268] Ibid.
[269] Meurens, N., et al. (2020). p. 28.
[270] Leye, E., et. al. (2021). p. 1.
[271] Ibid., p. 17.
[272] Meurens, N., et al. (2020). p. 42.
[273] Ibid.
[274] GREVIO (2020a). Para 40.
[275] Ibid.
[276] Leye, E., et. al. (2021). p. 5.
[277] Meurens, N., et al. (2020). p. 71.

provisions.[278] This fragmented approach has been criticised by GREVIO as failing to comprehensively address all manifestations of sexual harassment targeted by Article 40 IC.[279]

3.3. Legal Framework of the European Union on Sexfakes

With regard to the EU context, I will consecutively discuss the relevant legislation in chronological order. First, the GDPR will be discussed, then the AVMSD and the DSA after which the recently adopted AIA will be reviewed. Lastly, the new Directive on Combating Violence against Women and Domestic Violence will be discussed.

Disclaimer: EU Charter on Fundamental Rights

The Charter of Fundamental Rights of the European Union[280] can be considered the EU version of the CoE's ECHR. The Charter was adopted as a non-binding text in the year 2000 and became binding in December 2009 along with the adoption of the Treaty of Lisbon.[281] Most of the Articles in the Charter are inspired by the ECHR.[282] Article 6(3) TEU explicitly recognises the main rights of the ECHR as general principles of EU law and as a guiding influence for the development of EU fundamental rights.[283] Moreover, under the 'conformity clause'[284] of Article 52(3) of the Charter, the Court of Justice of the European Union (CJEU) is committed to uphold the ECHR standards and to align its interpretation of corresponding rights under the Charter with the case law of the ECtHR.[285] The relevant rights discussed above in the section on the CoE are therefore also protected by the EU Charter and will not be discussed in this section.

[278] Meurens, N., et al. (2020). p. 71.
[279] Ibid; GREVIO (2020b), para. 199.
[280] European Parliament, the Council and the Commission, Charter of Fundamental Rights of the European Union, OJ C 202/02, 7.6.2016.
[281] FRA. 'EU Charter of Fundamental Rights'. Retrieved from: https://fra.europa.eu/en/eu-charter#:~:text=The%20Charter%20of%20Fundamental%20Rights%20of%20the%20European%20Union%20(CFREU,the%20scope%20of%20EU%20law. On 2 July 2024.
[282] In the EU Charter the freedom of expression is cited in Article 11, the right to private and family life in Article 7 and the Prohibition of torture and inhuman or degrading treatment or punishment in Article 4.
[283] Article 6(3) TEU.
[284] Gil Carlos Rodríguez Iglesias (2002). 'Speech on the occasion of the opening of the judicial year, 31 Jan 2002' in *Annual Report 2001*, ECHR, Registry of the European Court of Human Rights Strasbourg, 2002, 3.
[285] Skouris, V. (2012). p. 2; Jacobs, F. (2008); Laffranque, J. (2012), p. 127; Kargopoulos, A. I. (2015).

3.3.1. General Data Protection Regulation

The GDPR entered into force on 24 May 2016 and applies since 25 May 2018.[286] It offers one of the strictest privacy and security legislation in the world.[287] Interestingly, the obligations set forth in the GDPR are also applicable to organisations outside of the EU, if they target or collect data from individuals within the EU.[288] One of the main aims of the Regulation is to protect fundamental rights and freedoms of individuals especially their right to personal data protection.[289]

Article 4(1) GDPR states that:

> *"'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".*

Yildrim and Aydinli argue that since deepfakes are fake in nature, applying data protection rules might be irrelevant, as the content does not pertain to a real person.[290] I do not agree with this argument. The latter statement could be true in cases where a deepfake is made without referencing an actual source, as the person depicted is non-existent and cannot be traced back to an identifiable person.[291] However, in the cases of non-consensual sexfakes, the victim's data and images are essential in creating the final deepfake, making it easy to trace back the video or image to herself.[292] I would argue that under the definition of Article 4 GDPR, deepfakes

---

[286] EC, 'Data Protection in the EU'. Retrieved from: https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en#:~:text=The%20General%20Data%20Protection%20Regulation%20(GDPR),-Regulation%20(EU)%202016&text=A%20single%20law%20will%20also,applies%20since%2025%20May%202018. On 6 July 2024.
[287] Wolford, B., 'What is GDPR, the EU's new data protection law?' Retrieved from: https://gdpr.eu/what-is-gdpr/ on 29 April 2024.
[288] Ibid.
[289] Article 1 GDPR.
[290] Yildrim, B., & Aydinli, C. (2019), 'Deepfake: An Assessment From The Perspective Of Data Protection Rules' Retrieved from: https://www.mondaq.com/turkey/privacy-protection/863064/deepfake-an-assessment-from-the-perspective-of-data-protection-rules on 5 July 2024.
[291] Okolie, C. (2023). p. 10.
[292] Ibid.

could be considered as personal data as the format in which the data is kept is irrelevant as long as the information can be traced to a real person.[293]

The GDPR applies to the 'processing' of personal data. According to Article 4(2) GDPR, processing entails "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaption or alteration". It can be concluded from this definition that the creation of a sexfake would qualify as data processing under Article 4(2) GDPR as it entails modifying the original data. I would therefore argue that a sexfake would fall within the scope of the Regulation.

The processing of personal data subsequently requires a legal basis.[294] In the context of deepfakes, two out of six possible legal grounds could be applicable.[295] Namely, 'informed consent' and 'legitimate interests'.[296] A satirical deepfake of a famous person could be protected under the right to freedom of expression and would therefore form a legitimate interest in processing an individual's personal data in the deepfake.[297] In the case of non-consensual sexfakes however I would argue that there is no legitimate interest in the creation and dissemination of these videos.

If legitimate interest does not apply, the deepfake creator could possibly process the personal data on the basis that the individual who is depicted consented to the creation of a deepfake image/video.[298] In the case of a sexfake, both the original person who is engaging in sexual activities should be asked for consent as the person whose face is copied onto the content of the other person's body.[299] This as both of their personal data are processed.[300] This legal ground does not work for my thesis either as I am writing about non-consensual sexfakes, where no prior consent was being asked by the perpetrator.

[293] Okolie, C. (2023). p. 10.
[294] Huijstee, M.V. et al. (2021)., p. 39.
[295] Ibid.
[296] Ibid.
[297] Ibid.
[298] Ibid.
[299] Ibid.
[300] Ibid.

The next question is what individual rights are protected under the GDPR. The GDPR puts strict conditions on the processing of personal data which means that not only the non-consensual creation and dissemination of sexfakes will fall under the restrictions prescribed by the GDPR, but also the training of software that is used for making sexfakes.[301] A Data Protection Impact Assessment (DPIA) has to be made when a service is operated that facilitates the creation of a deepfake.[302] An individual can also derive a range of different rights from the GDPR which are, among others, the right of access, the right to data portability, the right to restrict processing, the right to rectification, the right to object and the right to be forgotten.[303] Under the latter right, a data subject can ask the personal data controller such as the creator, publisher, or third-party search engines, directing users to sites containing deepfake pornography to delete, delink, or restrict further access to the content (Article 17 GDPR).[304] The RTBF is an important tool for victims of non-consensual sexfakes to have their rights protected.

First, the RTBF can limit the online damage faced by victims as their material will not be found any more in online searches after deletion.[305] This measure will assist victims in maintaining their financial stability and allows them to stabilise their lives more quickly after the incident as they will not have to be afraid that they will lose their jobs because of the reputational damage.[306]

The RTBF will also help victims to take back the victim's control over the situation. The non-consensual sharing of sexfakes is a huge violation of someone's autonomy. Removing non-consensual sexfakes empowers victims to regain control over their personal information by allowing them to decide who can access it. Consequently, the RTBF not only restores control to victims but also enhances their autonomy and privacy online.[307]

Lastly, the RTBF can serve as a legal tool to protect victims from further abuse and provides them with means to combat their abusers.[308] Where previously it was very difficult to ask online

[301] Huijstee, M.V. et al. (2021)., p. 39.
[302] Article 35 GDPR; Huijstee, M.V. et al. (2021)., p. 39.
[303] Okolie, C. (2023). p. 10.
[304] Ibid.
[305] Nguyen, T.N.A. (2022). p. 62-64.
[306] Ibid.
[307] Ibid.
[308] Ibid.

platforms to remove their sexfakes, the GDPR now mandates data controllers hosting abuse materials on their platforms to have proper systems and guidelines in place to help victims of deepfake pornography.[309] In practice, the RTBF has positively influenced platforms' policies on image-based sexual abuse.[310] For example, Google, after losing a case before the CJEU in 2014[311], now has to comply with European users' requests for the removal of "inadequate, irrelevant or excessive" data.[312] Afterwards, Google changed its policy to remove "non-consensual explicit or intimate personal images", "involuntary fake pornography", and "content about me on sites with exploitative removal practices" from its search results.[313] Similarly, Facebook has policies in place since 2017 to protect victims of image-based sexual abuse.[314] The platform prohibits sharing or threatening to share sextortion content, revenge pornography or non-consensual intimate images.[315] Facebook includes a report link on every piece of content so it can be reported easily.[316] The impact of the GDPR even surpasses EU borders; in the case of Facebook, the reporting and removal rules do not only apply in the EU, but everywhere in the world.[317]

Even though the RTBF is an important tool for addressing sexfakes, a key limitation arises from the lack of a clear definition of 'erasure' when applying the RTBF.[318] Technically, 'erasure' means completely wiping out data so it cannot be recovered, whereas 'delete' means removing data from the system's storage, making it recoverable until it is overwritten with new information.[319] The GDPR does not specify how data should be erased, leaving digital platforms to decide on their own what they will do.[320] For example, Google uses the term 'remove' for RTBF requests but removed URLs can still be accessed with different keywords.[321] Similarly, Facebook asserts that data subjects have the right to 'erase' their data under the GDPR but provides no further details, making it unclear whether the data is genuinely gone or just

[309] Nguyen, T.N.A. (2022). p. 62-64.
[310] Ibid.
[311] CJEU *Google Spain SL, Google Inc. / Agencia Española de Protección de Datos (AEPD), Mario Costeja González*, C-131/12.
[312] Ibid.
[313] Nguyen, T.N.A. (2022). p. 62-64.
[314] Ibid.
[315] Facebook, 'Not Without My Consent: A guide to reporting and removing intimate images shared without your consent'. Retrieved from: https://about.fb.com/wp-content/uploads/2017/03/not-without-my-consent.pdf on 28 June 2024.
[316] Ibid.
[317] Nguyen, T. N. A. (2022). p. 62-64.
[318] Ibid., p. 65.
[319] Gutmann & Warner (2019). p. 2; Nguyen, T.N.A. (2022). p. 65.
[320] Nguyen, T.N.A. (2022). p. 65.
[321] Ibid.

forgotten, potentially recoverable later.[322] This ambiguity in the definition of 'erasure' potentially undermines the effectiveness of the RTBF for victims of non-consensual sexfakes.

3.3.2. Audiovisual Media Services Directive

The AVMSD was revised and adopted in 2018 to address the expanding media landscape to include online video-sharing platforms.[323] It includes several guidelines aimed at preventing harm.[324] Article 6a(1) states that: "The most harmful content, such as gratuitous violence and pornography, shall be subject to the strictest measures".[325] The AMVSD does not provide a definition for 'gratuitous violence' except in Recital 20, which states that this content should not constitute a criminal offence.[326] The protection measures listed under Article 6a are general, such as age verification systems encryptions and effective parental control for the most harmful content.[327] This leaves a wide margin of appreciation for MS to impose stricter measures for various forms of harmful content.[328]

The AVMSD mandates that video-sharing platforms impose measures to prevent the dissemination of harmful content and ensure effective content moderation, including removal (Article 28b (3) AVMSD). These include specifying requirements in terms of services, user-friendly reporting mechanisms, and human moderation to detect and filter the most harmful content.[329]

However, the abovementioned rules only apply to protect minors. Only underage girls will therefore be protected by the AVMSD and not adult women. Furthermore, Article 6(a) does not prohibit the harmful content, only imposes 'strict measures'. In my opinion, these measures are inadequate to tackle non-consensual sexfakes are an age verification system or parental will not actually prevent the content from being created or distributed.

[322] Nguyen, T.N.A. (2022). p. 65.
[323] Huijstee, M.V. et al. (2021)., p. 42.
[324] Ibid.
[325] Article 6a(1) AVMSD
[326] EC & OSCE (2022). p. 44.
[327] Ibid.
[328] Ibid.
[329] Ibid.

3.3.3. Digital Services Act

In response to emerging national legal fragmentation, the EC announced the DSA to establish liability rules and content moderation obligations across the EU for digital platforms, services and products.[330] The DSA aims to clarify and expand a common set of responsibilities for online businesses providing services in the EU, regardless of the location of their headquarters.[331] Since 17 February 2024, these rules apply to *all* online platforms (irrespective of their size).[332] When the existence of illegal content is reported to the online platform, they should take effective measures, including the removal or disability of access to the illegal content (Article 6(1)(b), 9(1) DSA and recital 22 DSA). Illegal content is defined in Article 3(h) DSA as "any information that, in itself or in relation to an activity […] is not in compliance with Union law or the law of any MS which is in compliance with Union law, irrespective of the precise subject matter or nature of that law". The DSA therefore does not provide a list of what falls under illegal content. I would argue though that the sharing of non-consensual sexfakes would fall under the definition of illegal content of the DSA. In April 2024, the European Commission launched a new set of stringent obligations for adult entertainment platforms Pornhub, Stripchat and Xvideos. These platforms "must put in place mitigation measures to address risks linked to the dissemination of illegal content online, such as child sexual abuse material, and content affecting fundamental rights, such as the right to human dignity and private life in case of non-consensual sharing of intimate material online or deepfake pornography."[333] Furthermore, the new Directive on Combating Violence against Women - which I will extensively discuss in section 3.3.5. – stipulates that MS shall ensure the criminalisation of "producing, manipulating

[330] The DSA was adopted on 4 October 2022 by the European Council;
Huijstee, M.V. et al. (2021). p. 41.
[331] Council of the EU (2021). 'What is illegal offline should be illegal online: Council agrees position on the Digital Services Act'. Retrieved from: https://www.consilium.europa.eu/en/press/press-releases/2021/11/25/what-is-illegal-offline-should-be-illegal-online-council-agrees-on-position-on-the-digital-services-act/#:~:text=The%20rules%20set%20out%20under,should%20also%20be%20illegal%20online on 28 June 2024.
[332] European Commission (2024). 'Additional obligations for Very Large Online Platforms kick in for Pornhub, Stripchat and XVideos under the DSA'. Retrieved from : https://digital-strategy.ec.europa.eu/en/news/additional-obligations-very-large-online-platforms-kick-pornhub-stripchat-and-xvideos-under-dsa#:~:text=19%20April%202024-,Additional%20obligations%20for%20Very%20Large%20Online%20Platforms%20kick%20in%20for,Digital%20Services%20Act%20(DSA). On 29 April 2024.
[333] European Commission (2024). 'Additional obligations for Very Large Online Platforms kick in for Pornhub, Stripchat and XVideos under the DSA'. Retrieved from : https://digital-strategy.ec.europa.eu/en/news/additional-obligations-very-large-online-platforms-kick-pornhub-stripchat-and-xvideos-under-dsa#:~:text=19%20April%202024-,Additional%20obligations%20for%20Very%20Large%20Online%20Platforms%20kick%20in%20for,Digital%20Services%20Act%20(DSA). On 29 April 2024.

or altering and subsequently making accessible to the public, by means of ICT, images, videos or similar material making it appear as though a person is engaged in sexually explicit activities, without that person's consent, where such conduct is likely to cause serious harm to that person".[334] I would therefore conclude that non-consensual sexfakes fall under the definition of illegal content under the DSA, thereby mandating online platforms to take effective measures to remove non-consensual sexfakes from their platforms.[335] If an online platform does not take the relevant measures to remove the content, they can be subjected to severe fines. Article 52(3) DSA renders that MS shall set maximum fines for non-compliance with the DSA at 6% of the provider's annual global turnover from the previous year. For providing false, incomplete, or misleading information, failing to respond or correct such information, or obstructing inspections, maximum fines shall be 1% of the provider's or person's annual income or global turnover from the previous year.[336]

3.3.4. Artificial Intelligence Act

The newest addition to the EU legal framework is the EU AI Act.[337] The AIA establishes harmonised rules for the development, market placement and use of AI systems.[338] The fast developments within the field of AI made it necessary for the Commission to significantly change the AIA during the legislative process.[339] For example, generative and general-purpose AI was a technology which was not considered by the Commission during the initial drafting of the proposal.[340] However, the core principles, together with the AI Act's risk-based approach, have still been preserved.[341] In the next few years, the AIA will be further detailed and enhanced through secondary EU legislation.[342]

---

[334] Art. 7(b) Directive 2024/1385.
[335] See also: De Vido (2024). p. 3.
[336] Article 52(3) DSA.
[337] Published in the OJ on 12 July 2024, the AIA will enter into force 20 days after publication, which is 1 August 2024.
[338] Huijstee, M.V. et al. (2021)., p. 37.
[339] Sidley (2024), 'EU Formally Adopts World's First AI Law'. Retrieved from: https://datamatters.sidley.com/2024/03/21/eu-formally-adopts-worlds-first-ai-law/#:~:text=On%20March%2013%2C%202024%2C%20the,in%20favor%20of%20the%20legislation on 29 April 2024.
[340] Ibid.
[341] Ibid.
[342] Ibid.

The AIA takes a risk-based approach to possible AI threats. It categorises AI applications based on their risk levels into three categories. Firstly, applications and systems that pose an unacceptable risk, like social scoring systems similar to those in China, are prohibited (Article 5 AIA).[343] The second category concerns high-risk applications, such as CV-scanning tools that can scan CVs to rank job candidates (Article 6 AIA).[344] These applications have to meet extensive risk management, maintain transparency, and involve human oversight.[345] The third category requires applications with a limited risk to fulfil transparency requirements, ensuring that humans know that they are interacting with AI (Article 50 AIA). Deepfakes fall under this category (Article 50(4) AIA). Lastly, applications that fall under the category of minimal risk, like an AI spam filter, do not have any restrictions or mandatory obligations under the AIA.[346] In the figure below the risk-based approach is depicted as a pyramid.

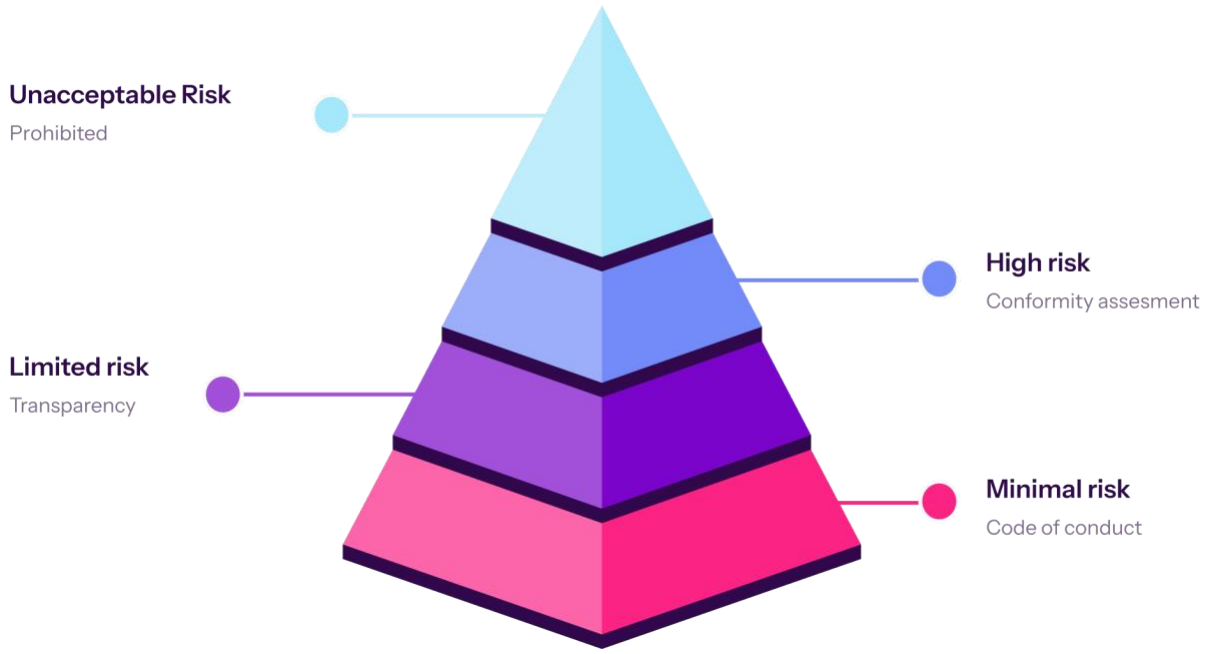


Figure 4: Risk-based approach in the AIA. Image pyramid retrieved from: https://www.credo.ai/eu-ai-act.

[343] Article 5 AIA, recital 31 AIA; EU Artificial Intelligence Act, 'The EU Artificial Intelligence Act: Up-to-date developments and analyses of the EU AI Act' retrieved from: https://artificialintelligenceact.eu/ on 29 April 2024.
[344] Article 6 AIA; Article 4 Annex III, high-risk AI systems referred to in Article 6(2) AIA; Benedek, W. (2023). Digital Human Rights and Artificial Intelligence. *Union UL Sch. Rev.*, *14*, 227, p. 236
[345] Article 6 AIA; Article 4 Annex III, high-risk AI systems referred to in Article 6(2) AIA; Benedek, W. (2023). Digital Human Rights and Artificial Intelligence. *Union UL Sch. Rev.*, *14*, 227, p. 236
[346] Trail, 'EU AI Act: How risk is classified. Retrieved from: https://www.trail-ml.com/blog/eu-ai-act-how-risk-is-classified on 28 June 2024.

Circling back to the obligation set for the creators of deepfakes, Article 52(3) AIA states that they shall disclose that the content has been artificially generated or manipulated.[347] However, Article 52(3) also states that this labelling requirement does not apply "where the use is authorised by law to detect, prevent, investigate prosecute criminal offences or it is necessary for the exercise of the right to freedom of expression and the right to freedom of the arts and sciences". In my opinion, non-consensual sexfakes do not fall under the exception of freedom of expression. The same reasoning applies here as in section 3.2.3. on the ECHR and Article 10.

If an operator or notified body does not comply with the requirements under Article 52(3), they can be subjected to an administrative fine of up to 15.000.000 Euros, or if the offender is an undertaking, receive a fine up to 3% of their total worldwide annual turnover for the preceding financial year, whichever is higher.[348] Despite, this hefty fine, the AIA is, in my opinion, not an effective instrument to fight against the creation and dissemination of sexfakes.

Firstly, while the transparency and labelling requirement mandates disclosure, it does nothing to prevent the creation or distribution of this harmful content. For victims, the primary concern is the complete removal of sexfakes, not merely their labelling as artificial. This approach misses the mark by prioritizing transparency over the more important need for content erasure and prevention.

Secondly, the scope of the transparency and labelling requirement for deepfakes under the AIA is still unclear,[349] as the AIA lacks specific guidelines for disclosure by creators.[350] Article 96(1)(d) AIA stipulates that the Commission shall develop guidelines on the implementation of the AIA, including the practical implementation of transparency obligations laid down in Article 50 AIA. However, due to the recent adoption of the AIA, the Commission has not provided these guidelines yet.

[347] Article 52(3) AIA.
[348] Article 99(4)(g) AIA.
[349] Huijstee, M.V. et al. (2021)., p. 39.
[350] Ibid.

Thirdly, the impact of the transparency requirement in the AIA on sexfakes made of famous individuals has to be seen.[351] Labelling sexfakes is unlikely to deter perpetrators, as the demand for such content does not hinge on its authenticity, and viewers often assume the content is fake anyway.[352]

Fourthly, the AIA neglects to acknowledge the disproportionate impact of sexfakes on women and girls. Preambular recitals 56 and 57 of the AIA only briefly mention that AI systems used in education and recruitment may perpetuate historical patterns of discrimination, for example against women.[353] When it comes to deepfakes, however, nothing is said about the disproportionate effect on women.

At the time of writing this thesis, the AIA has not entered into force yet – it will do so on 1 August 2024. It will be interesting to see what the impact of the transparency and labelling obligation will have on sexfake content, but as reasoned above, the measures proposed by the AIA appear inadequate to address this issue comprehensively. Consequently, I anticipate that the positive impact of the AIA will be limited.

3.3.5. Directive on Combating Violence against Women and Domestic Violence

The Directive on Combating Violence against Women and Domestic Violence (Directive 2024/1385) came into force in June 2024[354], providing measures on the non-consensual distribution of intimate and manipulated images.[355] When drafting the Directive, the Commission was inspired by the CoE's Istanbul Convention. At the time of the proposal, the EU had signed but not yet ratified the Istanbul Convention. This, as certain MS blocked the ratification of the Convention due to concerns regarding the reference to the notion of gender in its text.[356] The Commission's legislative proposal for Directive 2024/1385 was therefore

[351] Grady, P. (2023). 'EU proposals will fail to curb nonconsensual deepfake porn*'. Retrieved* from: *https://datainnovation.org/2023/01/eu-proposals-will-fail-to-curb-nonconsensual-deepfake-porn/#:~:text=Existing%20and%20proposed%20laws%20will,material%20without%20the%20individual's%20consent*. on 17 April 2024.
[352] Ibid.
[353] Preamble 56 and 57 AIA.
[354] Directive 2024/1385 has been published in the OJ on 24 May 2024, entering into force on 13 June 2024.
[355] Kuzmicz, M. M. (2023). p. 334.
[356] Rigotti, C. & McGlynn, C. (2022). p. 468-469.

designed to reach the same goal as the Istanbul Convention. Since then, however, the EU has acceded to the IC.[357]

Article 5(b) concerns the non-consensual sharing of intimate or manipulated material. MSs shall ensure the criminalisation of "producing, manipulating or altering and subsequently making accessible to the public, by means of ICT, images, videos or similar material making it appear as though a person is engaged in sexually explicit activities, without that person's consent, where such conduct is likely to cause serious harm to that person".[358] With this Article, the Commission has tried to criminalise the creation and dissemination of non-consensual sexfakes. Article 5(c) also provides for the criminalisation of sextortion as it states that it should be a criminal offence to "threaten to engage in the conduct referred to in point (b) in order to coerce another person to do, acquiesce or refrain from a certain act". In terms of sanctions, Article 10(4) provides that MSs should ensure that the criminal offences in Article 5 must be punishable by a maximum term of imprisonment of at least one year.

Regarding the mental element, Article 5 explicitly solely refers to the intentional distribution of sexfakes without consent, thereby not limiting the scope of the Article to cases where the abuser had a specific motive for his actions, like causing distress.[359] This inclusive approach thereby rejects the prevailing idea that abusers create and disseminate 'revenge porn' in order to harm the victim.[360] It emphasises the primary issue of non-consent, asserting that criminal penalties should apply irrespective of the perpetrator's motives.[361] This approach also simplifies legal proceedings, as evidence shows that a higher threshold would deter criminal justice staff from pursuing prosecutions.[362]

Article 23(1) of the new Directive mandates MS to "take the necessary measures to ensure that online publicly accessible material" as referred to in Article 5 is "promptly removed or that access thereto is disabled." MS must ensure that these actions are taken following transparent procedures and subjected to suitable safeguards, in particular to guarantee that the actions taken

[357] The EU has acceded to the Istanbul Convention in 2023 (Council of the EU (2023), 'Combatting violence against women: Council adopts decision about EU's accession to Istanbul Convention' retrieved from: https://www.consilium.europa.eu/en/press/press-releases/2023/06/01/combatting-violence-against-women-council-adopts-decision-about-eu-s-accession-to-istanbul-convention/ on 29 April 2024).
[358] Art. 7(b) Directive 2024/1385.
[359] Rigotti, C. & McGlynn, C. (2022). p. 472.
[360] Ibid.
[361] Ibid.
[362] Ibid.

are limited to what is necessary and proportionate, while also considering the rights and interests of all relevant Parties involved, including their fundamental rights as outlined in the Fundamental Rights Charter.[363]

While the Directive marks significant progress in acknowledging the gender-specific harm caused by non-consensual sexfakes, it still leaves several critical gaps unaddressed by the EU legislators.

Firstly, the Directive is based on Article 83(1) TFEU which grants competence to set minimum standards and define criminal offences "in the areas of particularly serious crime with a cross-border dimension resulting from the nature or impact of such offences or from a special need to combat them on a common basis".[364] Crimes that fall within the scope of Article 83 are among others 'computer crime' and 'sexual exploitation of women and children'.[365] The Directive's Articles on forced marriage and FGM are based on the 'sexual exploitation of women and children'[366], but strangely, the Articles on the non-consensual sharing of intimate images and sextortion are classified under 'computer crime' defined as any criminal offence "against or intrinsically linked to the use of information and communication technologies".[367] Even though it is applaudable that the Directive's focus on cyber violence recognises the significant harm caused to women online, legally categorising it as a computer crime risks perpetuating distinctions between online and offline behaviours, which do not align with women's actual experiences.[368] The abuse women face when confronted with sexfakes is not restricted to the online world as we have seen in Chapter 2. It also impacts their daily lives and wellbeing. Categorising sexfakes as a computer crime signals that the abuse faced is taken less seriously.[369]

Secondly, Article 5(b) of Directive 2024/1385 only applies to manipulated material where the victim is 'engaged in sexually explicit activities'.[370] The interpretation of this term may vary significantly, leading to definitional ambiguities.[371] It seems that images made with nudification

[363] De Vido (2024). p. 4.
[364] Article 83(1) TFEU.
[365] Rigotti, C. & McGlynn, C. (2022). p. 469.
[366] See: Chapter 2, Articles 3-9 Directive 2024/1385.
[367] Rigotti, C. & McGlynn, C. (2022). p. 469.
[368] Ibid.
[369] Ibid., p. 471.
[370] Ibid., p. 473.
[371] Ibid.

apps like 'ClothOff' will not fall within the scope of Article 5(b) as nudity is not per se sexual, and the images which are sexual often do not show people 'engaged' in sexual activities.[372] This legislative decision is especially confusing because Article 5(a) does refer to "intimate images, or videos or other material depicting sexual activities", thereby also encompassing nude images. In my view, this distinction seems arbitrary.

Thirdly, the Directive only covers the distribution of non-consensual sexfakes and not the creation.[373] Restricting the Article solely to distribution fails to capture the full extent of abuse and the experiences of female victims.[374] Victims do not perceive the abuse in distinct categories of 'creating' and distributing' but instead as a continuous experience.[375] Differentiating legal categories and laws based on specific actions does not align with victims' experiences.[376] Additionally, the non-consensual creation of sexfakes is in itself a severe violation of privacy and sexual autonomy and should therefore be criminalised accordingly.[377]

Fourthly, the wording of Article 5(c) Directive 2024/1385, sextortion, should be criticised. Article 5(c) lacks comprehensiveness as it only covers actions intended 'to coerce another person to do, acquiesce or refrain from a certain act'.[378] Even though this clause criminalises certain forms of blackmailing and threats in order to demand money or more pictures, it does not cover threats made solely to cause distress to the victim.[379] For instance, an ex-partner might threaten to distribute sexfakes simply to cause distress instead of meaning to coerce their ex-partner to do a certain act.[380] Similarly, other perpetrators may issue threats with the intent to exert power or control over the victim.[381] Thus, although the provision's inclusion of threats is a positive step, its limitation to certain types of threats fails to fully protect victims and leaves significant gaps in their protection.[382]

[372] Rigotti, C. & McGlynn, C. (2022). p. 474.
[373] Ibid.
[374] Ibid.
[375] Ibid.
[376] Ibid.
[377] Ibid.
[378] Ibid, p. 475.
[379] Ibid.
[380] Ibid.
[381] Ibid.
[382] Ibid.

Fifthly, the penalty linked with the crimes in Article 5, is criticised by some as being inadequate.[383] Article 10(4) stipulates that the criminal offences outlined in Article 5 must carry a maximum penalty of at least one year of imprisonment.[384] The European Economic and Social Committee (EESC) in its assessment of the Directive's proposal has criticised this maximum penalty, arguing that the sanction should match the minimum penalty for cyber-stalking, which was a maximum of two years' imprisonment.[385] However, in the final version of the Directive, the text has been amended to reduce the penalty for cyber-stalking instead, making it equivalent to the penalty for the distribution of sexfakes.[386] It is unfortunate to see that the final version has lowered the penalty for cyberstalking to the same level as that of manipulated materials instead of upping the penalty for Article 5(b).

Sixthly, and lastly, Article 5(b) is only applicable when the manipulated material is 'likely to cause serious harm to that person'. In the proposal for the Directive, this line was not included. It simply said that distribution of the manipulated material without that person's consent was punishable as a criminal offence. During the trialogue, this amendment was added and it seems to be a political compromise between the Parliament and the Council. This amendment has heightened the threshold for prosecution and it will therefore make it more difficult to prosecute the creator who made the sexfake.[387] Article 5 of the Directive neglects to recognise the fact that sexfakes in themselves are inherently harmful and should therefore be criminalised.[388]

3.4. Conclusion

In the previous sections, I have analysed and criticised the current European framework on non-consensual sexually explicit deepfakes, by looking both at the CoE and EU regulatory instruments.

In the EU context, I have discussed the GDPR, AVMSD, DSA, AIA, and Directive 2024/1385. With regards to the former four I would like to make some general remarks, as besides the gaps

[383] Rigotti, C. & McGlynn, C. (2022). p. 475; see also: EESC Committee (2022). Para. 3.18.
[384] Art 10(4) Directive 2024/1385; Rigotti, C. & McGlynn, C. (2022). p. 475.
[385] EESC (2022). para 3.18.
[386] Art 10(4) Directive 2024/1385.
[387] Belloni, P. (2024), 'The Proposal for a Directive on Gender-Based Violence'. Retrieved from: https://www.medialaws.eu/the-proposal-for-a-directive-on-gender-based-violence/ on 18 June 2024.
[388] De Vido (2024). p. 4.

already pinpointed in section 3.3., there are also some overarching gaps to be identified. The main problem with these instruments is that they do not provide enough tools for the enforcement of the law. Currently, only the perpetrator is held responsible for the dissemination of the content.[389] However, many perpetrators remain anonymous in order to evade law enforcement. When even the platform provider cannot identify them, no enforcement is possible.[390] Furthermore, sexfake victims often find themselves powerless to take effective action. Once a deepfake video starts circulating online, the victim usually loses control over it.[391] The victims are responsible for tracking down their sexfakes online, contacting data controllers and convincing them to take action against the abusive materials.[392] This is especially problematic because a lot of victims do not even know that their sexfakes are disseminated online and only find out about them after the sexfakes have been circulating for a long time already.[393] The time between distributing, discovering, reporting, and taking down the sexfake is therefore significant.[394] It is therefore questionable whether it is appropriate to ask victims to hold the offenders accountable.[395] By placing the burden on victims, who are mostly vulnerable women, the emotional distress is bigger than it should be.[396] Especially considering that the success of reporting and taking down of sexfake is not guaranteed.[397]

An issue which arises in both frameworks is the non-existence of jurisprudence on the topic of sexfakes (or even deepfakes for that matter). Even though I have argued that f.e. Articles 8 and 10 ECHR apply in the case of non-consensual sexfakes, legal uncertainty persists regarding its definitive applicability in this context. Especially in the EU context the CJEU has not had an opportunity yet to interpret the provisions of both Directive 2024/1385 and the AIA, as the Directive has only come into force in June 2024, and the AIA will enter into force in August 2024. Therefore, no conclusion can be drawn yet on the impact of these two instruments on a practical level.

However, to conclude, the most promising and comprehensive instruments so far are the CoE Istanbul Convention read in connection with GREVIO recommendation No. 1, and EU

[389] Huijstee, M.V. et al. (2021)., p. 50.
[390] Ibid.
[391] Ibid.
[392] Nguyen, T.N.A. (2022). p. 67.
[393] Ibid.
[394] Ibid.
[395] Huijstee, M.V. et al. (2021)., p. 50.
[396] Nguyen, T.N.A. (2022). p. 67.
[397] Ibid.

Directive 2014/1385. These instruments specifically address sexfakes as well as acknowledge the gendered dimension of the harm experienced. When comparing the frameworks of the EU and the CoE though, I believe that the EU framework is more effectively equipped to address sexfakes specifically. The EU's ability to directly harmonise legislation across all its MSs once adopted gives it a strong advantage. Furthermore, the DSA and GDPR make it possible for victims to remove their sexfakes from online platforms, thereby complementing the rules set down in Directive 2024/1385.

On the other hand, the EU's competence is first and foremost socio-economic in nature.[398] Therefore, the CoE normative human rights framework definitely has an advantage when thinking about sexfakes as a human rights violation. Furthermore, GREVIO Recommendation No. 1 clarified that *producing* as well as *procuring* sexfakes fall under the Istanbul Convention, contrary to Directive 2024/1385, where only the distribution of sexfakes is prohibited. Ideally, therefore, these two frameworks should be seen as complementary rather than competing with each other.

Now that I have discussed the EU and CoE framework for 'regular' sexfakes, I will continue my research in the next chapter by discussing the regulatory instruments' applicability for sexfakes in the Metaverse. Are the EU and CoE ready for this upcoming challenge or will the very recently adopted regulatory framework soon be outdated again?

[398] Wouters, J. et al. (2020). p. 11.

Chapter 4: Regulating Sexfakes in the Metaverse

4.1. Introduction

Section 4.2. will discuss the ways in which the Metaverse differs from the 'normal' online world. First, the issue of identity and personhood in the Metaverse will be discussed after which I will elaborate on the ownership question of avatars. Afterwards, it will be investigated whether or not we can speak of harm in the Metaverse in cases of sexfakes, after which the issues relating to perpetrator identification will be considered. Lastly, the question of jurisdiction in the Metaverse will be examined and the right to be forgotten.

In 4.3., I shall discuss who should be responsible for making sure that women are protected in the Metaverse from unwanted sexfakes: is self-regulation of the online platforms enough or should the (EU/CoE) legislator jump in?

In section 4.4., I shall take a step back and look at possible solutions from a broader perspective. Instead of looking at only regulatory solutions, the importance of deepfake detection models and Metaverse literacy will be explored.

Lastly, a conclusion will be drawn in section 4.5.

4.2. Online World vs Metaverse: Same System in a Different Font?

4.2.1. Identity

The first issue I want to discuss is that of the notion of identity in the Metaverse. Cornu defines 'identity' as 'what makes an individual himself and not another; by extension, what allows to recognise and distinguish him from others; [...] the set of characteristics that allow to identify him or her'.[399] Identity is dynamic, adaptable, complex and fluid.[400]

[399] Cornu, G. (2007). *Vocabulaire juridique*. 8th edition, Paris, P.U.F., p. 463.; Levallois-Barth, C. (2020). p. 1.
[400] World Economic Forum (2024). Metaverse Identity: Defining the Self in a Blended Reality. Insight Report March 2024. 1-48, p. 8.

Personal identity pertains to the unique ways individuals define themselves (the 'I').[401] It involves those traits that one believes define them as a person or make them who they are, including for example musical taste, clothing style or team affiliation.[402] Your personal identity is not set though, it typically changes over time: as a teen, you like pop music better whereas as an adult you may gravitate to indie rock.[403]

Social identity on the other hand refers to how people categorise themselves in relation to their group memberships (the 'we').[404] These categorisations are often assigned or inherent, such as political affiliations, possessions or club membership.[405]

The Metaverse stretches and reimagines the concept of personal and social identity, providing users with new ways for self-exploration and expression.[406] In the Metaverse one can change their appearance, adopting different forms like animals or mythical creatures.[407] The user has full control over their appearance.[408] The Metaverse allows users to bring their social identity to the Metaverse, to establish trust and authority, or to leave it behind, by embracing pseudonymity.[409] This adaptability within the Metaverse accommodates both those seeking continuity and those desiring reinvention in their digital interactions.[410]

However, while the Metaverse offers users the ability to embody virtually anything, including inanimate objects like lamps or a book, Hackl et al. suggest that most users are unlikely to adopt

---

401 Oswego. Social Identity. Retrieved from: https://ww1.oswego.edu/diversity/day-3-social-identity#:~:text=Personal%20Identity%20markers%20are%20often,something%20we%20are%20born%20into. On 1 July 2024.

402 Stanford Encyclopaedia of Philosophy. Personal identity. Retrieved from: https://plato.stanford.edu/entries/identity-personal/ on 1 July 2024; Oswego. Social Identity. Retrieved from: https://ww1.oswego.edu/diversity/day-3-social-identity#:~:text=Personal%20Identity%20markers%20are%20often,something%20we%20are%20born%20into. On 1 July 2024.

403 Shoemaker, P. (2022), What Is Digital Identity in the Metaverse? Retrieved from: https://www.identity.com/identity-in-the-Metaverse/#Identity_in_the_Metaverse on 11 June 2024

404 Oswego. Social Identity. Retrieved from: https://ww1.oswego.edu/diversity/day-3-social-identity#:~:text=Personal%20Identity%20markers%20are%20often,something%20we%20are%20born%20into. On 1 July 2024.

405 Shoemaker, P. (2022), What Is Digital Identity in the Metaverse? Retrieved from: https://www.identity.com/identity-in-the-Metaverse/#Identity_in_the_Metaverse on 11 June 2024.

406 Ibid.

407 Ibid.

408 Ibid.

409 Ibid.

410 Ibid.

such different forms.[411] Instead, those who feel constrained by their physical bodies, or face discrimination, may be more inclined to explore avatars that transcend their real-world limitations, and present themselves in the Metaverse as a 'normal person'.[412] That way they will be able to participate and integrate better in the Metaverse.[413] For example, a woman may choose to represent herself as a male avatar in order to avoid harassment or to increase her employability.[414] However, while this strategy might offer short-term benefits, it could ultimately undermine societal diversity.[415] According to Rigotti, adopting conformist avatars to escape discrimination or sexual abuse is not an effective solution to addressing real-world social inequalities.[416] Instead, this approach may actually compromise individual autonomy and self-expression.[417]

4.2.2. (Legal) Personality

The starting point of the scope of human rights is that the addressee is a human being.[418] This human being is a natural person who is a person by virtue of being born.[419] This is why I could go to court if my right to physical integrity has been violated, but a lamp that was smashed to pieces cannot.

For example, when we look at the wording of the Istanbul Convention and Directive 2024/1385, it is clear that only a human being is protected, and not an avatar in the Metaverse. This can be derived from the word 'person' used in the relevant Articles.

However, this does not automatically mean that if an avatar is compromised in the Metaverse by another avatar, legal redress is not possible.[420] Non-natural or artificial persons can still have a 'legal personality'. To qualify as a legal person, one has to be recognised by the legal system

[411] Hackl, C., et. al. (2022).
[412] Ibid.
[413] Burrows, G. (2022).; Rigotti, C., & Malgieri, G. (2023).
[414] Rigotti, C., & Malgieri, G. (2023).
[415] Ibid., p. 17.
[416] Ibid., p. 21.
[417] Ibid.
[418] Anže Mediževec (2024), 'AI Technology as a Legal Entity and its Protection from Discrimination', panel discussion at the Global Conference on AI and Human Rights taking place from 13 to 14 June 2024.
[419] Dyschkant, A. (2015).
[420] Kolen, E. et al. (2024). p. 6.

as a subject of the law, capable of bearing both responsibilities and rights.[421] Or as Salmond states:

> *"So far as legal theory is concerned, a person is any being whom the law regards as capable of rights and duties. Any being that is so capable is a person, whether a human being or not, and no being that is not so capable is a person, even though he be a man. Persons are the substances of which rights and duties are the attributes."*[422]

According to this definition, the notion of a 'legal person' is socially constructed; recognition as a legal person only requires a legal authority, such as the legislature, to decide that.[423] This is how previously corporations have become recognised as legal entities.[424]

According to Cheong, avatars should get legal personality when they are built by AI in a way that allows them to learn from their human users, make independent decisions, execute contracts and overlook others within the Metaverse.[425] In practice, legal personality could then be established through registration, with each individual entitled to only one avatar in the Metaverse.[426]

Until the time that we grant avatars legal personality, we still have to make sure that offences in the Metaverse can be prosecuted. Shoemaker therefore suggests that in the meanwhile, an individual's real-world legal identity would bind them in the digital realm as well.[427] Accountability for the offences committed within the Metaverse will stay with the human being who created the avatar, and not the avatar itself.

4.2.3. Ownership

The question arises whether we 'own' our avatar in the Metaverse – i.e. whether we own the data that comprises the avatar. Ownership is especially relevant if a victim wants to find legal redress for the creation and dissemination of Metaverse sexfakes. If the victim does not own

---

[421] Gunkel, D. J., & Wales, J. J. (2021). p. 474.
[422] Salmond, J.W. (1902). p. 334-335.
[423] Gunkel, D. J., & Wales, J. J. (2021). p. 474.
[424] Ibid.. 475.
[425] Cheong, B. C. (2022). p. 471-472.; Kolen, E. et al. (2024). p. 6.
[426] Ibid.
[427] Shoemaker, P. (2022), What Is Digital Identity in the Metaverse? Retrieved from: https://www.identity.com/identity-in-the-Metaverse/#Identity_in_the_Metaverse on 11 June 2024.

her avatar, she probably will not be able to start legal proceedings if the rights of her avatar are infringed either.

Do we have an implicit right to our data? It seems intuitive to believe that the data should be yours, as the avatars are created 'out of the fruits of our own labour'.[428]

However, as often, things are not as straightforward as they seem. Traditional legal definitions of property, struggle to apply to data due to issues with establishing exclusive possession.[429] Typically, for something to be considered owned, it must be finite and specific, allowing the owner to restrict others' access to that right.[430] Data, however, is fundamentally different. It's potentially infinite and often created through the involvement of multiple parties.[431] As a result, many legal experts argue that data ownership is not a viable concept.[432]

This is why the EU approaches personal data through privacy rights rather than ownership. Users grant permission to companies and governments to access their data based on consent.[433] EU citizens can, because of among others the GDPR, decline services that collect data (opt-out) and request the deletion of their personal information from company databases (RTBF).[434] However, the idea that individuals can truly opt out of data-reliant services in today's digital world is arguably unrealistic.[435] The significant power imbalance between data controllers and average citizens makes it difficult to view privacy as a genuine choice.[436] Many users inadvertently sign away their data rights by agreeing to complex platform privacy policies without fully understanding the implications.[437] A recent example is that of Meta's new privacy policy change on 26 June 2024.[438] Users who agreed with this new policy would consent to the use of all their publicly shared posts, images, image captions, comments and Stories on Facebook and Instagram to train Meta's AI products.[439] According to Meta, this is allowed under the GDPR,

---

[428] Mack, O. (2019), 'Who Owns Your Digital Twin? Not You—and Here's Why That's a Massive Problem'. Retrieved from: https://www.newsweek.com/who-owns-your-digital-twin-not-you-heres-why-thats-massive-problem-opinion-1451991 on 5 July 2024.; Dolan, L. (2022). p. 44.

[429] Rozynek, M. (2022). 'Me, myself and my avatar: Data ownership in virtual worlds'. Retrieved from: https://atelier.net/insights/me-myself-and-my-avatar-data-ownership-in-virtual on 12 July 2024.

[430] Ibid.

[431] Ibid.

[432] Ibid.

[433] Ibid.

[434] Ibid.

[435] Ibid.

[436] Ibid.

[437] Ibid.

[438] Macmahon, L. (2024), 'Plans to use Facebook and Instagram posts to train AI criticised'. Retrieved from: https://www.bbc.com/news/articles/cw99n3qjeyjo on 11 June 2024.

[439] Ibid.

as users will have the possibility to opt-out by visiting the Privacy Policy page within their Facebook and Instagram apps, accessible through the Settings and About screens, and selecting the Right to Object checkbox (which I have done so myself and it was definitely not an easy thing to do).[440]

Users outside of the EU are out of luck: There is no opt-out option available.[441] Max Schrems, famous from the CJEU cases Schrems I[442] and Schrems II[443], has already lodged complaints through his NGO NOYB ('None of your Business') in 11 EU MS and is urging authorities to initiate an emergency measure to stop this policy change immediately.[444]

> Max Schrems: *"Meta is basically saying that it can 'use any data from any source for any purpose and make it available to anyone in the world' as long as it does so via 'AI technology'. This clearly contradicts the GDPR. 'AI technology' is an incredibly broad term and Meta does not say what purposes it will use the data for. It could therefore be a simple chatbot, extremely aggressive personalised advertising or even a killer drone. Meta even says that it can make the data available to any 'third party'."*[445]

Your Metaverse avatar is built out of data, data that Meta will probably try to claim as their own, just like they are doing now with your photos on Instagram and Facebook. It will be interesting to see what the national privacy authorities will decide regarding the policy changes of Meta, as it will set a precedent of how far Meta will be able to stretch the boundaries of the GDPR. If they are able to collect and use all your data on online platforms for 'AI training purposes', what will hold them back in the Metaverse?

4.2.4. Harm

In order for a victim of a Metaverse sexfake to press charges, she first has to prove that there was harm.[446] Currently, it remains unclear whether a sexfake involving an avatar would be

---

[440] Speed, R. (2024), 'Meta faces multiple complaints in Europe over plans to train AI on user data'. Retrieved from: https://www.theregister.com/2024/06/06/meta_ai_complaints/ on 2 July 2024.
[441] Ibid.
[442] CJEU *Maximillian Schrems/ Data Protection Commissioner*, C-362/14.
[443] CJEU *Data Protection Commissioner / Facebook Ireland Ltd, Maximillian Schrems*, C-311/18,
[444] None of Your Business (NOYB) (2024). noyb fordert 11 Behörden auf, Metas Missbrauch persönlicher Daten für KI zu stoppen. Retrieved from: https://noyb.eu/de/noyb-urges-11-dpas-immediately-stop-metas-abuse-personal-data-ai on 2 July 2024.
[445] Ibid.
[446] Cheong, B. C. (2022). Kolen, E. et al. (2024). p. 7.

considered harmful in the same way as an online sexfake created of a real person. I would argue though that the impact experienced in the Metaverse by sexfakes does pass the harm threshold. I would even go as far as to say that Metaverse sexfakes could lead to even more harm than an online sexfake. In the following paragraph, I will explain why.

A crucial aspect of the Metaverse is that it is immersive: the whole goal of the Metaverse is that experiences in the virtual world feel intensely real, just as they would in the real world.[447] This means that compared to 'regular' sexfakes, the psychological effect of such virtual experiences can be even more severe on victims when seeing them.[448] It feels like they are genuinely raped in real life when engaging in the Metaverse.

The argument that no harm is done to the real-world person as the sexfake is made of the avatar does not work here either. Studies indicate that individuals develop a strong emotional bond with their avatars[449], or as Rigotti says: "[when a user enters a virtual environment, the virtual world becomes their world, their avatar becomes their body.[450] This deep identification with the avatar suggests that sexual cyber violence will also be processed by the brain similarly to real-world incidents, causing the same psychological harm and traumatic response.[451]

Because of the foregoing, no significant distinction should be made between virtual and physical life as it is important to view sexual misconduct as part of the broader continuum of violence that women face in all aspects of their lives[452], including the Metaverse.[453] We can see how the cascading effect of sexfakes also applies here in the Metaverse, as it not only affects the individual victim but also society as a whole.[454]

[447] Marr, B. (2024). 'The Metaverse And Its Dark Side: Confronting The Reality Of Virtual Rape' retrieved from: https://www.forbes.com/sites/bernardmarr/2024/01/16/the-Metaverse-and-its-dark-side-confronting-the-reality-of-virtual-rape/ on 18 March 2024.
[448] Ibid.
[449] Dolan, L. (2022). p. 40.
[450] Rigotti, C., & Malgieri, G. (2023). p. 23.
[451] Dolan, L. (2022). p. 44.; Wiederhold, B. K. (2022). p. 479.
[452] Kelly, L. (2013).
[453] Rigotti, C., & Malgieri, G. (2023). p. 23.
[454] Ibid.

4.2.5. Perpetrator

Finding the perpetrator who created and disseminated non-consensual sexfakes is going to be more difficult in the Metaverse as well. Currently, more than 50% of sexfakes victims are in the unknown of the identity of the perpetrator[455], and this number will only grow in the Metaverse as they will have more ways to hide their identity – by their choice of avatar as well as the anonymity provided for by blockchain technology.[456] There is therefore a growing feeling among abusers that they can exploit the anonymity provided by the Metaverse without having any consequences.[457]

This is not a new problem. Anonymity has always been strongly linked to the internet – even going as far as being perceived as 'the cornerstone' of the internet - as it was a place where people could freely speak their minds without being afraid of negative consequences in the real world.[458] When it comes to the Metaverse, the European Citizens' Virtual Worlds Panel (which was assigned by the EU Commission in 2023 to give recommendations on the values and actions needed to create attractive and fair European virtual worlds) stressed in their Recommendation number 19 that: "There should be a regulation at the EU-level on when you need to show your identity and when you can be anonymous in the digital world. When we talk about entertainment, leisure, or research, it should be possible to be anonymous."[459] However, this anonymity cannot be unlimited. In certain situations, it will be necessary to authenticate yourself with a digital identification.[460] The Citizens' Virtual Worlds Panel gives as an example the need for identification in cases of transferring money, using government services or when buying specific goods where a license or an age limit is requested.[461] In my opinion, users should only be able to participate in the Metaverse when they identify themselves first when signing up.[462] Pseudonymity might be the solution here: users would fully disclose their identity with the Metaverse platform they are trying to enter but when they engage in that Metaverse

[455] Laffier, J. & Rehman, A. (2023)., p. 5.
[456] Gadekallu, T. R., et al. (2022).
[457] Cheong, B. C. (2022).
[458] Davenport, D. (2002). p. 33; Kolen, E. et al. (2024). p. 8.
[459] European Commission (2023), 'European Citizens' Virtual Worlds Panel: A new phase of citizen engagement' retrieved from: https://citizens.ec.europa.eu/virtual-worlds-panel_en#:~:text=The%20expected%20outcome%20of%20the,an%20initiative%20on%20the%20topic. On 12 June 2024.
[460] Ibid.
[461] Ibid.
[462] Kolen, E. et al (2024). p. 8.

they use their pseudonym.[463] Pseudonymity is not a new concept and has been used for hundreds of years by artists who produce work under a different name for a multitude of reasons.[464] In my opinion, pseudonymity could offer the anonymity needed when engaging with others in the Metaverse, but will not exempt individuals from prosecution in cases of non-consensual sexfakes.[465]

4.2.6. Jurisdiction and Statehood

Sixthly, questions of jurisdiction arise. In the physical world, jurisdiction is typically linked to the location of the crime, the victim or the perpetrator; however, the whole point of the Metaverse is that it transcends physical borders and is completely decentralised.[466]

A solution would be to recognise the Metaverse as a 'Network State', which is:

> *"a highly aligned online community with a capacity for collective action that crowdfunds territory around the world and eventually gains diplomatic recognition from pre-existing States."*[467]

A Network State differs from a Nation State, as the latter is connected to land whereas the former is related to a person's mind.[468] In other words, while the Nation State system begins with a map of the globe and allocates each piece of land to a specific State, the Network State system starts with the 7+ billion people in the world and connects each individual to one or more networks.[469]

While interesting, the Network State would still come with their own problems, as we do not know who would be in charge of creating these new rules. Could the current international community bind the avatars in the Metaverse with their rules or should the avatars themselves come up with these rules? The latter seems more difficult than the former, but the former is also

[463] Kolen, E. et al (2024). p. 8; Shoemaker, P. (2022), What Is Digital Identity in the Metaverse? Retrieved from: https://www.identity.com/identity-in-the-Metaverse/#Identity_in_the_Metaverse on 11 June 2024.
[464] Ibid.
[465] Kolen, E. et al (2024). p. 8.
[466] Yılmaz, H. K. E. (2024). p. 53-54.
[467] Srinivasan, B. (2022), p. 9.
[468] Ibid.
[469] Ibid.

difficult as creating a universal set of rules and regulations for the Metaverse is challenging because different States have varying interpretations of what would constitute criminal behaviour.[470] A separate law for the Metaverse would also isolate the event as something only relevant for the Metaverse, whereas I just discussed in the previous section that the harm and consequences are very much felt by the female victim in the real world. Lastly, by making the Metaverse a Network State, it would assume that the Metaverse is something completely distinct from the real world. However, Zuckerberg envisions the Metaverse as something hybrid, not as something completely separate. If the Metaverse develops in the former way, the Network State may not be a possible solution.

4.2.7. Right to be Forgotten

Lastly, I shortly want to touch upon the right to be forgotten, as protected under Article 17 GDPR and 8 ECHR. The Metaverse is completely decentralised through the use of blockchain. Blockchain holds numerous identical copies of a database on various computers distributed throughout a network.[471] Once it is on the blockchain, it is there forever.[472] Reports of Europol[473] and the European Parliament[474], articles from Rigotti[475] and Tatar[476] have argued

[470] Tariq, S., Abuadbba, A., & Moore, K. (2023, July). Deepfake in the Metaverse: security implications for virtual gaming, meetings, and offices. In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes* (pp. 16-19), p. 18.

[471] Rodeck, D., & Curry, B. (2022). 'What is blockchain.' Retrieved from: https://www.forbes.com/advisor/in/investing/cryptocurrency/what-is-blockchain/ on 5 July 2024.

[472] Europol (2022)., p. 17.

[473] Europol (2022)., p. 17:
"It is possible to send anyone an NFT or message on the blockchain, but once it is on the blockchain there is no way for anyone to remove it; this will mean any harassment may indefinitely show up if people look into your blockchain, blocking any way out of that abuse."

[474] European Parliament (2019) Blockchain and the General Data Protection Regulation Can distributed ledgers be squared with European data protection law? STOA.
p II: "It is the tension between the right to erasure (the 'Right to be forgotten') and blockchains that has probably been discussed most in recent *years. Indeed, blockchains are usually deliberately designed to render the (unilateral) modification of data difficult or impossible. This, of course, is hard to reconcile with the GDPR's requirements that personal data must be amended (under Article 16 GDPR) and erased (under Article 17 GDPR) in specific circumstances."*
p 75 : "Many have stressed the difficulty of applying the right to erasure to blockchains. Deleting data from DLT is burdensome as these networks are often purposefully designed to make the unilateral modification of data hard, which in turn is supposed to generate trust in the network by guaranteeing data integrity. […]Indeed, even if there would be a means of ensuring compliance from a technical perspective, it may be organisationally difficult to get all nodes to implement related changes on their own copy of the database (particularly in public and permissionless blockchains). "

[475] Rigotti, C., & Malgieri, G. (2023). p. 11:
"The main obstacle for such people will be difficulty in understanding the complex functionality of blockchain, which will impair a user's autonomy and informational self-determination – for example, the user might find it difficult to exercise their Right to be forgotten on blockchain technology."

[476] Tatar, U., Gokce, Y., & Nussbaum, B. (2020).:

that erasure of content in the Metaverse will therefore not be possible. Sexfakes will not be able to be removed and they will indefinitely show up in your blockchain.[477]

How can the right to be forgotten be exercised in the Metaverse if it is not possible to erase sexfakes from the blockchain?

As part of my research, I talked to both a computer scientist[478] and a legal scholar[479] to ask more about the difficulties that arise with the RTBF in the Metaverse. Garcia, a computer scientist, has argued that it is technically possible to remove content from the blockchain. However, Rigotti, a legal scholar, refuted this statement by explaining that even though RTBF could be exercised in theory, it would not be possible in practice. This as the erasure is both time and effort-consuming, thereby putting an unfeasible burden on the Metaverse platforms to enforce.

Until an actual victim of a sexfake in the Metaverse seeks redress under Article 17 GDPR or 8 ECHR, the practical applicability of the Right to be Forgotten in this context remains uncertain. However, the divergent perspectives of technological and legal experts on this issue highlight a critical challenge in regulating emerging technologies. Lawmakers are tasked with crafting regulations for systems they may not fully comprehend from a technical standpoint. The contrasting views of Garcia and Rigotti exemplify this gap: while one argues for the technical feasibility of content removal from the blockchain, the other emphasises the practical impossibility due to the resource-intensive nature of the process. Therefore, in my opinion, interdisciplinary collaboration and communication are crucial. Without a comprehensive understanding of both the technological capabilities and legal implications, there is a risk of creating regulations that are either technologically infeasible or legally ineffective.

"[*Article 17] stands in contrast with one of the fundamentals of blockchain technology. […] It is not possible to remove data from a blockchain ledger without "breaking the chain." It would otherwise negate one of the very basic principles upon which the blockchain technology is based, namely irreversibility/immutability. […] A natural outcome of such a design is, any attempt to change or manipulate data stored in a block would distort the whole blockchain consistency. This means, when a company using the blockchain technology fulfils the request of data subjects who exercise their Right to be forgotten, they would do so at the expense of blockchain consistency, which would likely be detrimental to reliability and customer trust."*

[477] Ibid.

[478] Nuno Garcia, PhD. Associate Professor with Habilitation at the Computer Science Department at UBI and Invited Associate Professor at the Universidade Lusófona de Humanidades e Tecnologias. Interview on 20 May 2024 on Zoom.

[479] Carlotta Rigotti PhD. Post-doc researcher at eLaw – Center for Law and Digital Technologies, Leiden University. Interview on 18 June 2024, on Microsoft Teams.

4.3. Who Should Regulate the Metaverse?

In the previous section, I have discussed multiple possible issues that will arise in the Metaverse in the future. I am aware that it is not possible yet to answer these questions but in the following section I will at least try to answer who should be responsible for making an effort to regulate it. I will first discuss the possibility of self-regulation by Metaverse platform providers after which I will discuss the role of the European regulators.

4.3.1. Self-Regulation

As technological advancements by companies often outpace regulatory frameworks in setting standards and enforcement, companies are often urged to voluntarily take up a proactive approach in setting standards themselves and create their own community guidelines.[480] Self-regulation involves companies voluntarily identifying, assessing and mitigating potential risks associated with their activities in the Metaverse.[481]

It is not possible to assess the self-regulatory efforts of all Metaverse companies, but I would like to take Meta as an example to see what has already been done when it comes to self-regulation. Meta announced a $50m (€46m) investment programme in 2021 to ensure the Metaverse is developed responsibly.[482] In its 'Code of Conduct for Virtual Experiences' Meta explains that creators and administrators are primarily responsible for addressing issues within their governed experiences.[483] If issues stay unresolved by creators or administrators, developers must adhere to this code of conduct by restricting features or removing users from the app.[484] Meta will only step in when there are persistent or severe issues that cannot be resolved by either the administrators or the developers.[485]

[480] Benedek, W. (2023). p. 229-230.
[481] Benjamins, R., Rubio Viñuela, Y., & Alonso, C. (2023). p. 15-16.
[482] Meta (2021). 'Building the Metaverse responsibly'. Retrieved from: https://about.fb.com/news/2021/09/building-the-Metaverse-responsibly/ on 2 July 2024.
[483] Meta (2024). Code of Conduct for Virtual Experiences. Retrieved from: https://www.meta.com/en-gb/help/quest/articles/accounts/privacy-information-and-settings/code-of-conduct-for-virtual-experiences/ on 2 July 2024.
[484] Ibid.
[485] Ibid.

With regards to the conduct monitored, Meta considers that behaviours such as: "[…], Pretending to be another person or entity, stealing someone's identity, or creating or using fake accounts, […], [promoting] anything that's illegal, abusive or could lead to physical harm, such as: sexualising, exploiting or abusing minors, […], any form of non-consensual intimate activity, including sharing intimate images of others without consent", are against their values.[486]. Developers, creators and administrators within the Metaverse can also establish their own rules that exceed the Code of Conduct, which users also have to adhere to.[487]

Meta clearly tries to make the creators and administrators responsible for the bulk of the regulation necessary. Only as a last resort will Meta step in as a platform.

Meta also seems to focus more on protecting children than adult women from sexual abuse in the Metaverse. There is a separate page dedicated to the safety and privacy of teens in the Metaverse.[488] Even though it is good that there is a specific focus on teenagers, who in general are more vulnerable to sexual abuse, Meta's guidelines do not consider the gendered aspect of this vulnerability.[489] As discussed, teenage girls are disproportionally affected by cyber violence and sexual abuse in the offline and online world. Meta should also consider this when making protective guidelines.

Meta's code of conduct is a step in the right direction but definitely not sufficient yet. Their guidelines all fit on one page and so far, no efficient reporting mechanisms are in place.[490] Current users report challenges in identifying the speakers, noting that their usernames are not easily traceable and that it can be difficult for new users to understand how the reporting mechanism works.[491] It is also not clear what the capabilities of administrators are when they signal illegal behaviour.[492]

---

[486] Meta (2024). Code of Conduct for Virtual Experiences. Retrieved from: https://www.meta.com/en-gb/help/quest/articles/accounts/privacy-information-and-settings/code-of-conduct-for-virtual-experiences/ on 2 July 2024.
[487] Ibid.
[488] See: Meta (2024), Safety and privacy tools available for teens in Meta Horizon Worlds. Retrieved from https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/safety-and-privacy-tools-teens-horizon-worlds/ on 2 July 2024.
[489] Kolen, E. et al. (2024). p. 9.
[490] Mooij, A. & Tushuizen, J. (28 June 2024). Regulating the Virtual World as a new State . European Law Blog. Retrieved from: https://europeanlawblog.eu/2024/06/28/regulating-the-virtual-world-as-a-new-State / on 1 July 2024.
[491] Ibid.
[492] Ibid.

When it comes to transparency, Meta publishes transparency reports on how they enforce their policies.[493] However, no transparency reports have been published yet by Meta in which they give insights into their activities and enforcement in the Metaverse.

When it comes to oversight, it is not clear whether the Meta Oversight Board (OB) has a role to play in the Metaverse. Currently, Meta has an OB reviewing Facebook's, Instagram's, and Threads' content moderation decisions.[494] The OB is legally independent from Meta and provides binding decisions that Meta has to follow, as long as implementing these decisions does not violate the law.[495] The OB has announced two new cases for consideration in April 2024 which concern "explicit AI Images of Female Public Figures" (i.e. sexfakes).[496] They are currently assessing whether Meta's policies and its enforcement practices are effective at addressing explicit AI-generated imagery.[497] These cases align with the OB's 'Gender strategic priority' which they announced in 2022.[498] As part of this strategy, the OB specifically focuses on the gendered obstacles women face in exercising their rights to freedom of expression, such as gender-based violence and harassment, and examines the impact of gender-based distinctions in content policy.[499] The OB's focus on the gendered dimension of harassment online and its focus on sexfakes on Facebook and Instagram, also give hope for the Metaverse. However, as of July 2024, the mission statement on the OB's website states that: "The Oversight Board's mission is to improve how Meta treats people and communities around the world. We apply Facebook, Instagram and Threads' content standards in a way that protects freedom of expression and other global human rights standards.[500] Based on its wording, the OB does not seem to have competence in the Metaverse. This absence of competence could diminish the OB's significance, particularly if Meta continues to prioritise the Metaverse. However, if the Metaverse will be mainly integrated within existing platforms like Facebook, Instagram, or

[493] Meta. Transparency reports'. retrieved from: https://transparency.meta.com/reports/ on 3 July 2024.
[494] Wong, D., & Floridi, L. (2023). p. 262.
[495] Meta, 'Just the Facts on the Oversight Board'. Retrieved from: https://about.meta.com/actions/oversight-board-facts/ on 8 July 20204.
[496] OB (2024). Oversight Board Announces Two New Cases On Explicit Ai Images Of Female Public Figures. Retrieved from: https://www.oversightboard.com/news/oversight-board-announces-two-new-cases-on-explicit-ai-images-of-female-public-figures/ on 8 July 2024.
[497] Ibid.
[498] OB (2022). 'Oversight Board Announces Seven Strategic Priorities', retrieved from: https://www.oversightboard.com/news/543066014298093-oversight-board-announces-seven-strategic-priorities/ on 8 July 2024.
[499] Ibid.
[500] See: Oversight Board. Mission statement. Retrieved from: https://www.oversightboard.com/ on 1 July 2024.

Threads, Wong et al. argue that the OB should have jurisdiction in the Metaverse.[501] Only time will tell what will actually happen in the future.

4.3.2. Regulation

While community guidelines for content moderation can be a force for good and can have a genuine impact, they also face significant criticism.[502] These criticisms encompass concerns about biased motivations among administrators and influential community members who shape these guidelines, the inadequacy of vigilante actions in responding to spontaneous individual behaviours, and the potential risk for victims to be targeted in the future when they report such abuse.[503] Furthermore, it has been argued more generally that social media platforms often prioritise business interests over user rights, or better said: profit over people.[504]

That is why it is necessary to also have regulation from outside these platforms, from both the CoE as well as the EU. They have three options: 1) not doing anything[505], 2) amending the current instruments, or 3) drafting new instruments specifically targeting the Metaverse.

The first option seems to be the current approach of the EU and CoE.

On an EU level, the EC considered in its 'Communication on an EU initiative on Web 4.0 and Virtual Worlds' in July 2023 that it is unlikely that they would propose any new legislation on virtual worlds as the EU has a "robust, future-oriented legislative framework that already applies to several aspects of the development of virtual worlds and Web 4.0.".[506] The DSA, AIA, and GDPR "introduce a comprehensive system of accountability and obligations for online platforms", "establish horizontal rules for data-sharing and give users control over the data generated by their connected devices", and "tackle risks emerging from artificial intelligence (AI) and will promote innovation in trustworthy AI".[507]

To a certain degree, I agree with the Commission's argumentation.

[501] Wong, D., & Floridi, L. (2023). p. 276.
[502] Stavola, J., & Choi, K. S. (2023). p. 15.
[503] Ibid.
[504] Benedek, W. (2023). p. 229-232.
[505] Or at least, focusing on the effective implementation of their instruments by State Parties.
[506] European Commission (2023). p. 4-5.
[507] Ibid.

For example, the newly amended European Digital Identity (eIDAS Regulation)[508] would help with perpetrator identification. The Regulation provides for a framework for a European digital identity that is accessible to all EU citizens, residents, and businesses, through a European digital identity wallet (EDIW).[509] MS will provide businesses and citizens with a digital wallet linking their national digital identities to other personal attributes such as their driving license or bank account.[510] They can use their EDIW to access online services throughout the EU.[511] Even though this Regulation has just been adopted, this framework can have a big impact on how identification in the Metaverse will work. Within the Metaverse, the EDIW could serve as a vital tool for securing the digital identification of users, enhancing user control by offering features like pseudonymity, and facilitating 'zero-knowledge proof technology'.[512] The latter is a cryptographic method that allows a relying party to verify an EDIW's user statement without the need to reveal underlying data, thereby preserving the user's privacy.[513]

Furthermore, the jurisdictional question is maybe not as pressing as it may seem. Previously, scholars have already discussed this topic in the context of the internet. Some argued that as the internet has no physical borders, governance and jurisdiction should also reflect this borderless reality.[514] However, so far, this shift has not happened and we have seen that States and Courts find innovative solutions to circumvent the jurisdictional question. An example of this is the case of *Glawischnig-Piesczek v. Facebook.*[515] In this case, an Austrian politician was targeted by comments on Facebook which, according to Austrian law, were deemed illegal.[516] The CJEU was asked whether EU law permitted the national court to order Facebook to globally remove statements that had the same wording and/or conveyed equivalent content as the unlawful post

[508] Regulation (EU) 2024/1183 was published in the Official Journal of the European Union on 30 April 2024 and it entered into force on 20 May 2024. The eIDAS Regulation amends the previous Regulation (EU) No 910/2014 as regards establishing the European Digital Identity Framework.
[509] Press Release: European digital identity (eID): Council adopts legal framework on a secure and trustworthy digital wallet for all Europeans'. Retrieved from: https://www.consilium.europa.eu/en/press/press-releases/2024/03/26/european-digital-identity-eid-council-adopts-legal-framework-on-a-secure-and-trustworthy-digital-wallet-for-all-europeans/ on 10 June 2024.)
[510] Ibid.
[511] Ibid.
[512] Ramos, T. a.o. (2024), 'Navigating New Realities: The Impact of the revised eIDAS Regulation on the Metaverse and VLOPs' retrieved from: https://www.taylorwessing.com/en/insights-and-events/insights/2024/03/embracing-the-future-of-digital-identities on 10 June 2024.
[513] Ibid.
[514] Wanjiru, M. (2023). Beyond Borders: Understanding The Trends of Internet Jurisdiction. Retrieved from: https://paradigmhq.org/beyond-borders-understanding-the-trends-of-internet-jurisdiction/ on 3 July 2024.
[515] CJEU *Eva Glawischnig-Piesczek / Facebook,* C-18/18.
[516] Casolari, F., & Gatti, M. (2022). p. 20-21.

under Austrian law.[517] The CJEU held that it is possible for a court of a MS to: "[order] a host provider to remove information covered by the injunction or to block access to that information worldwide within the framework of the relevant international law".[518] This judgement seems to imply that the territorial scope of EU law could exceed EU borders. This would take the so-called 'Brussels effect' to an even higher level. The former entails the principle that when the EU comes up with new legislation, the rest of the world often follows suit by drafting new legislation inspired by that of EU law.[519] However after *Glawishnig*, this principle can be taken literally as Brussels' impact may now be directly exceeding EU borders, without the need for other jurisdictions to make up their own legislation.
I would therefore argue that, until the Metaverse is acknowledged globally as a Network State, and it is possible to become a legal resident of that Network State, an individual's legal identity in the physical world will continue to bind them in the digital realm.[520] Consequently, in my opinion, the individual should also be bound by the 'real world' legislation criminalising sexfakes.

However, gaps persist in the EU's response to these challenges, with various questions still awaiting clarification. For example, Article 5 of the Directive is only applicable "where such conduct is likely to cause serious harm to that person". Even though I have argued above that in my opinion the harm caused to victims can be considered very grave, the Commission has not shared its view on this topic yet in its Communication (this is because the Communication predates the adoption of the Directive), nor has the CJEU. Until that time, we will not know whether Directive 2024/1385 is also applicable in the Metaverse.

The same goes for the CoE. Even though the CoE report on the Metaverse published in June 2024 concluded that the assessment of whether existing frameworks are adequate or new ones are needed will require in-depth impact assessments,[521] it also stated that in principle all the relevant existing conventions and recommendations also apply in the Metaverse as "fundamental freedoms and human rights apply equally online and offline".[522] Accordingly, the

[517] Casolari, F., & Gatti, M. (2022). p. 20-21.
[518] CJEU *Eva Glawischnig-Piesczek v. Facebook*, C-18/18, para. 53.
[519] Dolan, L. (2022). p. 8.
[520] Shoemaker, P. (2022), What Is Digital Identity in the Metaverse? Retrieved from: https://www.identity.com/identity-in-the-Metaverse/#Identity_in_the_Metaverse on 11 June 2024.
[521] CoE & IEEE (2024). p. 59.
[522] EuroDIG 2024 – Balancing innovation and regulation. Launch Event for the Council of Europe and IEEE Joint Report on the Metaverse and its impact on Human Rights, Rule of Law, and Democracy, 17 June 2024.

problem does not lie with insufficient regulation, but with insufficient implementation by State Parties.[523] Even though I agree with the CoE stance, I also think that the current framework leaves room for further clarification on several fronts.

For example, would the ECtHR accept that a sexfake of an avatar constitutes a violation of the right to one's image under Article 8 ECHR? Even though the ECtHR has interpreted the word 'image' in a broad sense, it has to be seen whether we can talk about someone's image when they decide to create an avatar in the Metaverse which does not have any likeness or resemblance to the offline person; let alone a deepfake of that avatar in the Metaverse.
The same applies to the right to protect one's reputation. The attack on someone's private life has to "attain a certain level of seriousness and in a manner causing prejudice to personal enjoyment of the right to respect for private life".[524] To what extent can we even talk about an attack on an avatar's reputation? The same goes for the sufficient link between the applicant and the purported attack on their reputation.[525] Is there a sufficient link between the deepfake of the avatar and the creator of the avatar?

In general, the questions raised in 4.2. are not fully answered by the EU and CoE. Consequently, in this scenario, much power will be put in the hands of the Courts. They will be responsible for interpreting the instruments in a way which would/ or would not provide protection for Metaverse sexfake victims. Generally, the Courts interpret relevant provisions in a way that the function (purpose) of those provisions is included in the assessment.[526] It will be up to the Courts to decide whether or not a broad interpretation of the relevant provisions aligns with the objectives pursued by the CoE and EU with their regulations.

The second option would be to amend the existing framework, and the third, is to create a new Convention for the CoE or a new Directive or Regulation[527] for the EU. It falls outside of the scope of this thesis to extensively discuss the pros and cons of picking one option over the

[523] EuroDIG 2024 – Balancing innovation and regulation. Launch Event for the Council of Europe and IEEE Joint Report on the Metaverse and its impact on Human Rights, Rule of Law, and Democracy, 17 June 2024.
[524] ECtHR *Axel Springer AG v. Germany*, paras. 83-84.
[525] ECtHR *Putistin v. Ukraine*, para. 40.
[526] I.e. teleological interpretation; Zurek, T., & Araszkiewicz, M. (2013). p. 160.
[527] In this case the most logical choice would be a Directive based on Article 83 TFEU as this article concerns the harmonisation of criminal offences.

other.[528] No matter which route the CoE or EU will take though, it is important to clarify the following things.

First, we have to define what the Metaverse is. As said before, there is no universally agreed upon definition or consensus regarding the Metaverse.[529] The EU refers to it as 'virtual world' and 'immersive realities', whereas other international organisations like the CoE still use the term Metaverse (just like me).[530] Nor is there agreement over what the factual substance of the definition would be. The CoE even says in its report on the Metaverse that: "the report does not aim to provide a definition, but rather a description of the Metaverse".[531] This is rather odd in my opinion as coming from a legal perspective it is very important to always make sure that everyone has the same definition in mind when talking about a certain topic.

Second, the EU and CoE have to clarify who 'owns' their avatar in the Metaverse; the platform provider or the user. In my opinion, the latter should be the case. We cannot have meaningful participation in the Metaverse if we do not have full control over who we are. Nor can we find legal redress. This is especially important as the end goal of the Metaverse is to have one universal, cohesive, and interoperable 3D space that will integrate the numerous virtual worlds that exist today while at the same time being completely decentralised.[532] In this view of the Metaverse, you cannot have different avatars on different platforms, but you own one specific avatar which represents you in the Metaverse. This ownership should be reflected in the EU's and CoE's regulations.

Third, the possibility of conferring legal personality upon avatars should be considered. As said before, legal personality is socially constructed and only requires a legal authority to decide upon.[533] However, even though legal personality for avatars seems like a logical step to make,

[528] Amending the current CoE and EU instruments could be beneficial as it would build upon an already established legal structure and would integrate the challenges of the Metaverse into the broader context of the exsting framework. However, the current instrument could be too narrow to fully address the complexities of the Metaverse and some States might oppose reopening a Convention or Directive for amendments. Creating a new legal instrument on the other hand could be beneficial to address the unique challenges of the Metaverse but could also lead to extra fragmentation and would take time to draft and implement.
[529] Kitsara, I. (2024). 'The Metaverse and its impact on human rights, the rule of law and democracy'. Abridged version of the report for the Council of Europe. Retrieved from: https://rm.coe.int/the-Metaverse-impact-on-and-its-impact-on-human-rights-the-rule-of-law/1680ae6bce on 4 March 2024, p. 4.
[530] Ibid.
[531] CoE & IEEE (2024). p. 13.
[532] Ibid.
[533] Gunkel, D. J., & Wales, J. J. (2021). p. 474.

we should be wary of taking away responsibility from individuals for their actions and decisions in the Metaverse.[534] The perpetrator who creates and distributes a sexfake should not be exempt from prosecution in the real world just because his avatar has legal personality in the Metaverse. A dual system should apply where justice is served both in the real world and the Metaverse.

4.4. A Broader Perspective: Solutions Outside of the Regulatory Framework

4.4.1. Detection Models

On a practical level, it is important to create technical solutions to address the risks associated with sexfakes in the Metaverse and to evaluate how well current deepfake detection technologies can handle emerging challenges.[535] Detection models are important because they are a great tool in distinguishing between authentic and synthetic media sources.[536] Without detection models we would not even come to the question of how we should regulate sexfakes as we would not know in the first place that the material was fake.

Unfortunately, so far, no method exists yet in which we can detect sexfakes in the Metaverse that is already on the market.[537] However, multiple researchers have proposed different models that could potentially apply in the Metaverse.

Sensity[538] is like an anti-virus software but completely created to detect deepfakes, using the same deep learning processes as are used for creating deepfakes.[539] It notifies users by email upon detecting synthetic media fingerprints generated by AI.

Another study found that employing a single Res-NET-34 encoder in a hierarchical approach across three levels can effectively detect and provide detailed explanations of deepfakes, by

[534] Gunkel, D. J., & Wales, J. J. (2021). p. 476.
[535] Stavola, J., & Choi, K. S. (2023). p. 9.
[536] Ibid.
[537] Ibid., p. 16.
[538] See for the website of Sensity: https://sensity.ai/deepfakes-detection/
[539] Wu, H., et al. (2023). p. 5.

distinguishing fake data from real data.[540] The same researcher (Guarnera) proposed developing a model recognition system to trace AI images back to their generator or model owner.[541] This would be especially relevant in the Metaverse, where it could help attribute deepfake images to the specific user who created them.[542]

As we can see, the technology involved in deepfake detection models is rapidly evolving. Researchers seem to be aware of the risks associated with deepfakes in the Metaverse and are therefore focusing their efforts on developing effective detection systems. In my opinion, these detection models should be implemented into mainstream Metaverses like Roblox or Meta as soon as possible in order to assess where the strengths and weaknesses of these technologies lie.

4.4.2. Metaverse Literacy

Lastly, I want to stress that we should not underestimate the importance of what I call 'Metaverse literacy'. There is currently limited public awareness regarding the issue of sexfake creation in the Metaverse.[543] Given the rising popularity of the Metaverse, it is crucial to employ awareness techniques to prevent future illegal activities on the platform.[544] Metaverse users should get the tools and education to protect themselves from sexual abuse and report illegal sexfakes when necessary.

Some NGOs are working on public awareness already, with one of the examples being the 'Metaverse Safety Week' (MSW). The MSW is an annual awareness campaign established by X Reality Safety Intelligence (XRSI) to encourage a safe and enjoyable experience in virtual worlds.[545]

The European Citizens' Virtual Worlds Panel also recommended the EC to make a guideline on how to be a digital citizen, to be made by a panel of experts who come from various backgrounds.[546] This guideline should then be implemented by national governments in their

[540] Guarnera, L., et al. (2022). p. 3; Stavola, J., & Choi, K. S. (2023). p. 16.
[541] Ibid.
[542] Ibid.
[543] Stavola, J., & Choi, K. S. (2023). p. 16.
[544] Ibid.
[545] See: Metaverse Safety Week. Retrieved from: https://Metaversesafetyweek.org/about/ on 1 July 2024.
[546] Recommendation 16, European Citizen Virtual Worlds panel. Retrieved from: https://citizens.ec.europa.eu/virtual-worlds-panel_en on 1 July 2024.

education systems.[547] The Citizen's Panel also expects citizens to be active in the debate and follow the development of guidelines and policies relating to the Metaverse.[548] Citizens would be educated on how to avoid misinformation but also on their duty to give correct information and not to harm others in the Metaverse.[549]

4.5. Conclusion

In this chapter, I have shown that there are critical differences between traditional online spaces and the Metaverse. These are, among others, issues relating to personal and social identity, (legal) personality, avatar ownership, harm, perpetrator identification, jurisdiction and the right to be forgotten. These issues could be addressed by Metaverse platform providers through self-regulation and/or the EU and CoE through regulation. With regards to the former, I have used the example of Meta's Code of Conduct for Virtual experiences to show that despite the effort, limitations are still prevalent in addressing the full scope of potential harms of sexfake victims. It is therefore important that a strong regulatory framework exists. The EU's and the CoE's are both of the opinion that no new instruments for the Metaverse are necessary as the already existing framework fully applies in the Metaverse. In my analysis in 4.3.2. I have shown that this is only true to a certain extent. Perpetrator identification could be done more easily with the newly adopted European Digital Identity (eIDAS Regulation)[550] and the question of jurisdiction does not seem to be as pressing as at first glance, but still, multiple questions are left unanswered. The CoE's and EU's conservative approach could result in the underestimation of the unique nature of the Metaverse and its potential for gender-based cyber violence.
Lastly, I discussed how beyond legal and regulatory measures, technological solutions such as deepfake detection models tailored for the Metaverse environment are crucial. Additionally, promoting 'Metaverse literacy' through education and awareness campaigns is essential for empowering users to protect themselves and report illegal activities. It takes a concerted and collaborative effort among all stakeholders in the industry to establish and enforce standards and regulations that are capable of addressing the issue of sexfakes in the Metaverse.

---

[547] Recommendation 16, European Citizen Virtual Worlds panel. Retrieved from: https://citizens.ec.europa.eu/virtual-worlds-panel_en on 1 July 2024.
[548] Ibid.
[549] Ibid.
[550] Regulation (EU) 2024/1183 was published in the Official Journal of the European Union on 30 April 2024 and it entered into force on 20 May 2024..

5. General Conclusions

This thesis has examined the complex issue of non-consensual sexually explicit deepfakes (sexfakes) in the context of the emerging Metaverse, analysing the current European legal frameworks established by the CoE and the EU.

I started my thesis by giving an in-depth analysis of the case study of sexfakes. Deepfakes are images or recordings that have been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said.[551] While deepfakes can be used for good (in film-making, or research projects like 'deep empathy'), so far it has done more harm than good. This is exemplified by the massive creation and distribution of sexfakes. Of all deepfake content, 96% is sexually explicit[552], and within this category, women and girls make up 99% of the victims, making it an offence which almost exclusively targets women.[553] I argued that without acknowledging the gendered harm caused by sexfakes, we cannot fully grasp the severity of the harm either. Furthermore, I showed that the offence of sexfakes is not only a current issue we are facing on the internet but will continue to cause harm in the Metaverse. Even though currently no 'real' Metaverse exists, it has already been proven that once the Metaverse has fully emerged as a viable technology in the near future, it will also bring about sexfakes.[554] It is therefore crucial that a robust and effective regulatory framework is in place to tackle this issue.

Before, delving into the applicable regulatory framework in the Metaverse, it was important to first establish the applicable regulatory framework on 'regular' sexfakes in the online world. I consecutively discussed the CoE's Framework Convention on Artificial Intelligence and

[551] Definition deepfake. Retrieved from: https://www.merriam-webster.com/dictionary/deepfake on 8 April 2024.
[552] Openletter.net (2024), 'Deepfakes'. Retrieved from: https://openletter.net/l/disrupting-deepfakes on 22 June 2024.
[553] Tsalidis, A. (2024), 'Disrupting the Deepfake Pipeline in Europe' retrieved from: https://futureoflife.org/ai-policy/disrupting-the-deepfake-pipeline-in-europe/ on 9 April 2024; Openletter.net (2024), 'Deepfakes'. Retrieved from: https://openletter.net/l/disrupting-deepfakes on 22 June 2024; Compton, S. & Hamlyn, R. (2023), 'Opinion: The rise of deepfake pornography is devastating for women' retrieved from: https://edition.cnn.com/2023/10/29/opinions/deepfake-pornography-thriving-business-compton-hamlyn/index.html#:~:text=This%20practice%20is%20no%20longer,on%20victims%20can%20be%20devastating;; Tenbarge, K. 'Found through Google, bought with Visa and Mastercard: Inside the deepfake porn economy' retrieved from: https://www.nbcnews.com/tech/internet/deepfake-porn-ai-mr-deep-fake-economy-google-visa-mastercard-download-rcna75071 on 8 March 2024.
[554] Europol (2022)., p. 7-8.; Cubewealth (2023), 'what is the Metaverse: Origins, Platforms, Future, Warnings?' retrieved from: https://www.bankoncube.com/post/what-is-the-Metaverse on 18 March 2024.

Human Rights, Democracy and the Rule of Law, the Budapest Convention, the ECHR, and the Istanbul Convention, and the EU's GDPR, AVMSD, DSA, AIA, and Directive 2024/1385.

While the EU's Framework Convention on AI and the Budapest Convention have been proven to be insufficient to protect sexfake victims, I argued that sexfakes could fall under the scope of the ECHR, more specifically under Article 8 (privacy and physical integrity), Article 10 (freedom of expression), and Article 3 (freedom from degrading treatment). However, as the ECtHR has not come out with any judgment yet in which they discuss sexfake as a human rights violation under the ECHR, we do not know yet if it in fact applies. The Istanbul Convention read together with GREVIO Recommendation No. 1 is therefore more promising as they explicitly recognise sexfakes as a form of gender-based violence, urging for the criminalisation of both the production ánd procurement of non-consensual sexfakes.

The EU Directive 2024/1385 - the EU's equivalent of the Istanbul Convention - on the other hand only criminalises the distribution of sexfakes, not the creation. Furthermore, it requires the content to depict 'sexually explicit activities', potentially excluding nude images, and includes a 'serious harm' threshold, making it more difficult to prosecute. Still, it includes more possibilities for enforcement than the IC does. The Directive itself includes the mandatory maximum penalty of one year, and the GDPR, AVMSD and DSA require platform providers to remove this illegal content from their platforms. The new AIA is in this context not helpful though as it only includes a transparency obligation for deepfake creators to disclose the fact that the material is made synthetically.

In Chapter 3 I have therefore concluded that the EU approach may be more effective but the CoE's human rights-based framework provides a valuable complementary perspective.

Chapter 4 discussed how the current framework could be applied to the Metaverse. The Metaverse has its own unique challenges which affect the applicability of the existing regulatory framework.

The Metaverse allows for fluid personal and social identities, which may result in users, who feel constrained by their physical bodies, or face discrimination, to be inclined to present themselves in the Metaverse as a 'normal person' in order to avoid harassment. However, while this strategy might offer short-term benefits, it could ultimately undermine societal diversity.

The Metaverse also raises critical questions about the legal status of avatars and their connection to real-world individuals. Furthermore, ownership of avatars and associated data in the Metaverse has been proven to be ambiguous, potentially complicating legal recourse for victims of sexfakes. The immersive nature of the Metaverse may intensify the psychological impact of

sexfakes, possibly causing more severe trauma than traditional online content. This is compounded by the enhanced anonymity in the Metaverse, which could make it more difficult to identify and prosecute offenders. Furthermore, the decentralised, borderless nature of the Metaverse complicates traditional notions of legal jurisdiction, while its blockchain-based structure may challenge the implementation of the RTBF as protected under the GDPR and Article 8 ECHR.

The EU's and CoE's stance that no new instruments are necessary appears therefore overly optimistic and may underestimate the transformative potential of the Metaverse, especially in relation to gender-based cyber violence. Chapter 4 thus explored two main approaches to regulation: self-regulation by platform providers and regulation by European legislators. Using Meta as an example, I have demonstrated that while some efforts are being made through codes of conduct, they are currently insufficient to fully address the risks of sexfakes in the Metaverse. On the regulatory front, the EU and CoE could therefore decide to amend existing regulations in the field of data protection and gender-based violence or create a new regulatory instrument in which they specifically address the upcoming challenges in the Metaverse. Future regulations should prioritise first to clearly define the Metaverse, clarify avatar ownership, and consider granting legal personality to avatars while maintaining user responsibility. Beyond legal solutions, the chapter ended by emphasising the importance of developing detection models to identify sexfakes in the Metaverse and promoting 'Metaverse literacy' through education and awareness campaigns. A collaborative effort among all stakeholders is hereby necessary.

Circling back to Mary Shelley's Frankenstein, the foregoing analysis of sexfakes revealed the potential for new technologies to escape from our human control, resulting in severe harm. However, as the 'real Metaverse' has not come into full force yet, we are provided with the unique opportunity to intervene now, to ensure that these digital 'monsters' do not cause chaos in the Metaverse in the future. Where Dr Frankenstein failed to take effective action, the EU and CoE could. This is why the title of my thesis is "A Modern Tale of Frankenstein?". The question mark is deliberate, signifying that our story's ending remains unwritten. It will be up to us to write a different ending to our book so that we are not doomed to repeat Shelley's cautionary tale more than 200 years later. Both the EU and the CoE have an important role to play in this process, having the potential to become an inspiration for the rest of the world. By taking the lead, they could transform the Metaverse into a safe and inclusive digital space for all, for both women, men and everyone in between.

Bibliography

Literature

Açar, K. V. (2016). Sexual Extortion of Children in Cyberspace. International Journal of Cyber Criminology, 10(2).

Almenar, R. (2021). Cyber violence against Women and Girls: Gender-based Violence in the Digital Age and Future Challenges as a Consequence of Covid-19. Trento Student Law Review, 3(1), 167-230.

Benjamins, R., Rubio Viñuela, Y., & Alonso, C. (2023). Social and ethical challenges of the Metaverse: Opening the debate. *AI and Ethics*, *3*(3).

Benedek, W. (2023). Digital Human Rights and Artificial Intelligence. *Union UL Sch. Rev.*, *14*, 227.

Borghetto, E., & Mäder, L. (2014). EU law revisions and legislative drift. *European Union Politics*, *15*(2), 171-191.

Boyd, P. (2022). Fakes and Deepfakes: Balancing Privacy Rights in the Digital Age. *Ala. L. Rev.*, *74*, 517.

Burrows, G. (2022). *Your Life In The Metaverse: Everything you need to know about the virtual internet of tomorrow*. Gideon Burrows.

Casolari, F., & Gatti, M. (2022). The Application of EU Law Beyond Its Borders. *Cleer Working Papers*, 3, 1-246.

Cheong, B. C. (2022). Avatars in the Metaverse: potential legal issues and remedies. *International Cybersecurity Law Review*, *3*(2), 467-494.

Chowdhury, R. (2023). *Technology Facilitated Gender-Based Violence in an era of Generative AI*. UNESCO Publishing.

Committee on Artificial Intelligence (CAI) (2024). Draft Framework Convention on artificial intelligence, human rights, democracy and the rule of law: Draft Explanatory Report. CM(2024) 52- addprov.

Council of Europe & IEEE (2024). The Metaverse and its impact on human rights, the rule of law and democracy. Council of Europe Publishing.

Cybercrime Convention Committee (CAI), Working Group on cyberbullying and other forms of online violence, especially against women and children. Mapping Study on Cyber violence with recommendations adopted by the T-CY on 9 July 2018. Strasbourg, 9 July 2018. T-CY(2017)10.

Yang, C. Z., Ma, J., Wang, S., & Liew, A. W. C. (2020). Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. *IEEE Transactions on Information Forensics and Security*, *16*, 1841-1854.

Davenport, D. (2002). Anonymity on the Internet: why the price may be too high. *Communications of the ACM*, *45*(4), 33-35.

Delfino, R. A. (2019). Pornographic deepfakes: The case for federal criminalisation of revenge porn's next tragic act. *Fordham L. Rev.*, *88*, 887, 937.

De Vido (2024), Deep fake as AI-generated violence against women. *DEP - Deportate, esuli, profughe,* n. 19. 1-4.

Dolan, L. (2022). *The legal ramifications of virtual harms: a study into the human rights implications of a meta-led Metaverse*. Master dissertation European Master in Human Rights and Democratisation.

Dyschkant, A. (2015). Legal personhood: how we are getting it wrong. *U. Ill. L. Rev.*, 2075.

Eckert, S. (2018). Fighting for recognition: Online abuse of women bloggers in Germany, Switzerland, the United Kingdom, and the United States. New Media & Society, 20(4), 1282-1302.

European Commission (2023). Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions: An EU initiative on Web 4.0 and virtual worlds: a head start in the next technological transition. COM(2023) 442/final.

European Commission (2023). Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions: An EU initiative on Web 4.0 and virtual worlds: a head start in the next technological transition. COM(2023) 442/final.

European Court of Human Rights (ECtHR) (2016). Guide on Article 8 of the European Convention on Human Rights Right to respect for private and family life, 1-101.

European Economic and Social Committee (2022). Opinion on the Proposal for a Directive of the European Parliament and of the Council on combating violence against women and domestic violence. COM(2022) 105 final.

Europol (2022). Policing in the Metaverse: what law enforcement needs to know, an observatory report from the Europol Innovation Lab, Publications Office of the European Union. Luxembourg.

Fabbrini, F., & Celeste, E. (2020). The Right to be forgotten in the Digital Age: The Challenges of Data Protection Beyond Borders. *German Law Journal, 21*(S1), 55–65.

Fernandez, A. (2022). Regulating Deep Fakes in the Proposed AI Act. *En ligne], Law and Policy of the Media in a Comparative Perspective, billet de blog publié le*, *23*(03).

Gadekallu, T. R., et al. (2022). Blockchain for the Metaverse: A review. *arXiv preprint arXiv:2203.09738*.

Group of Experts on Action against Violence against Women and Domestic Violence (GREVIO) (2020a). 1st General Report On GREVIO's Activities, covering the period from June 2015 to May 2019. Council of Europe Publication Office.
Group of Experts on Action against Violence against Women and Domestic Violence (GREVIO) (2020b). Baseline Evaluation Report Italy. Council of Europe Publication Office.

Group of Experts on Action against Violence against Women and Domestic Violence (GREVIO) (2021). General Recommendation No. 1 on the digital dimension of violence against women. Council of Europe Publication Office.

Guarnera, L., Giudice, O., Nießner, M., & Battiato, S. (2022). On the Exploitation of Deepfake Model Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 61-70).

Gunkel, D. J., & Wales, J. J. (2021). Debate: what is personhood in the age of AI?. *AI & society*, *36*, 473-486.

Gutmann, A., & Warner, M. (2019). Fight to be forgotten: Exploring the efficacy of data erasure in popular operating systems. In Privacy Technologies and Policy: 7th Annual Privacy Forum, APF 2019, Rome, Italy, June 13–14, 2019, Proceedings 7 (pp. 45-58). Springer International Publishing.

Hackl, C., Lueth, D., & Bartolo, T. D. (2022). Navigating the Metaverse: A guide to limitless possibilities in a WEB 3.0 world. Wiley.

Huijstee, M. V., et al. (2021). Tackling deepfakes in European policy. Think Tank European Parliament. *The European Parliament*.

Jacobs, F. (2008). Pace, Committee on Legal Affairs and Human Rights, Doc 1533, 18/3/2008, 2008, Appendix, part B, IV;

Kargopoulos, A. I. (2015). ECHR and the CJEU: Competing, overlapping, or Supplementary Competences?. *Eucrim: the European Criminal Law Associations' fórum*, (3), 96-100.

Kelly, L. (2013). *Surviving sexual violence*. John Wiley & Sons.

Kolen, E. et al. (2024). Guardians of the METAlaxy: How can Meta users be protected from sexual abuse in the Metaverse? Essay for the Course 'AI and Human Rights' at Uni Graz. 1-18.

Kuzmicz, M. (2023). Naked in the Eyes of the Law: Criminal Law Perspective on Nudity in Chosen European Jurisdictions in the Context of Innovative Technologies. *European Journal of Crime, Criminal Law and Criminal Justice*, *31*(3-4), 325-345.

Kye, B., et. al. (2021). Educational applications of metaverse: possibilities and limitations. *Journal of educational evaluation for health professions*, *18*.

Jacobsen, B. N., & Simpson, J. (2023). The tensions of deepfakes. *Information, Communication & Society*, 1-15.

Laffier, J., & Rehman, A. (2023). Deepfakes and Harm to Women. *Journal of Digital Life and Learning*, *3*(1), 1-21.

Laffranque, J. (2012). Who has the last word on the protection of human rights in Europe. *Juridica Int'l*, *19*, 117.

Levallois-Barth, C. (2020). Calling for the recognition of a right to multiple digital identities.

Leye, E., D'Souza, H., & Meurens, N. (2021). The added value of and resistance to the Istanbul convention: a comparative study in 27 European Member States and Turkey. *Frontiers in Human Dynamics*, *3*, 697331.

Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, *7*(4), 415-423.

Mammadzada, I. (2021). Deepfakes and Freedom of Expression: European Perspective. MA thesis at Tallinn University of Technology.

McGlynn, C., Rackley, E., & Houghton, R. (2017). Beyond 'revenge porn': The continuum of image-based sexual abuse. *Feminist legal studies*, *25*, 25-46.

McGoldrick, D. (2013). The Limits of Freedom of Expression on and Social Networking Sites: A UK Perspective. *Human Rights Law Review*, *13*(1), 125-151.

Ng, D. T. K. (2022). What is the metaverse? Definitions, technologies and the community of inquiry. *Australasian Journal of Educational Technology*, *38*(4), 190-205.

Nguyen, T. N. A. (2022). European 'Right to be forgotten' as A Remedy For Image-Based Sexual Abuse: A Critical Review. *KnowEx Social Sciences*, *2*(01), 59-72.

N.-M. Aliman, L. Kester, Malicious design in aivr, falsehood and cybersecurity-oriented immersive defenses, in: 2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), IEEE, 2020.

Okolie, C. (2023). Artificial Intelligence-Altered Videos (Deepfakes), Image-Based Sexual Abuse, and Data Privacy Concerns. *Journal of International Women's Studies*, *25*(2), 11.
Popay, J., et al. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC methods programme Version, 1(1).

Onyshkevych, R. (2022). Issues in the Practice of Implementing the Istanbul Convention. *Teisinės minties šventė 2022. Studentų mokslinių straipsnių rinkinys/redaktorės Prof. dr. Eglė Bilevičiūtė, Greta Petkutė, dr. Kristina Kenstavičienė. ISSN 2783-6886*.

Rigotti, C., & Malgieri, G. (2023). Human vulnerability in the Metaverse. Brussels: Alliance for Universal Digital Rights & VULNERA.

Rigotti, C., & McGlynn, C. (2022). Towards an EU criminal law on violence against women: The ambitions and limitations of the Commission's proposal to criminalise image-based sexual abuse. *New Journal of European Criminal Law*, *13*(4), 452-477.

Salmond, J.W. (1902) *Jurisprudence, Or the Theory of the Law*. Stevens and Haynes, London.

Shelley, M. (2012). *Frankenstein*. Penguin Classics.

Skouris, V. (2012). 'Protection in the Union and the Charter of Fundamental Rights of the European Union', Contemporary issues relating to the protection of Fundamental Rights at the European Level, Hellenic National School of Judges, Thessaloniki.

Sloot, B., Wagensveld, Y., & Koops, B. J. (2021). Deepfakes. Tilburg Institute for Law, Technology, and Society.

Sorban, Kinga. (2021). The Evolution of Content-Related Offences and Their Investigation during the First 20 Years of the Cybercrime Convention. Hungarian Yearbook of International Law and European Law, 2021, 305-327.

Srinivasan, B. (2022). The network state: How to start a new country. *Kindle Edition*.

Stavola, J., & Choi, K. S. (2023). Victimisation by Deepfake in the Metaverse: Building a Practical Management Framework. *International Journal of Cybersecurity Intelligence & Cybercrime*, *6*(2), 2.

Tariq, S., Abuadbba, A., & Moore, K. (2023). Deepfake in the Metaverse: security implications for virtual gaming, meetings, and offices. In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, 16-19.

Tatar, U., Gokce, Y., & Nussbaum, B. (2020). Law versus technology: Blockchain, GDPR, and tough tradeoffs. Computer Law & Security Review, 38, 105454.

The Platform of Independent Expert Mechanisms on Discrimination and Violence against Women (EDVAW) (2022). *The digital dimension of violence against women as addressed by the seven mechanisms of the EDVAW Platform.* Publication office Council of Europe.

Van der Nagel, E. (2020). Verifying images: Deepfakes, control, and consent. *Porn Studies*, *7*(4), 424-429.

Van der Wilk, A. (2021). *Protecting women and girls from violence in the digital age*. Publication Office Council of Europe.

Velasco, C. (2022, May). Cybercrime and Artificial Intelligence. An overview of the work of international organisations on criminal justice and the international applicable instruments. In *ERA Forum* (Vol. 23, No. 1, pp. 109-126). Berlin/Heidelberg: Springer Berlin Heidelberg.

Veletsianos, G., et al. (2018). Women scholars' experiences with online harassment and abuse: Self-protection, resistance, acceptance, and self-blame. New Media & Society, 20(12), 4689-4708.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review*, *9*(11).

Wiederhold, B. K. (2022). Sexual harassment in the metaverse. *Cyberpsychology, Behavior, and Social Networking*, *25*(8), 479-480.

Wittes, B., Poplin, C., Jurecic, Q., & Spera, C. (2016). Sextortion: Cybersecurity, teenagers, and remote sexual assault. *Center for Technology Innovation at Brookings*, 1-47.

Wong, D., & Floridi, L. (2023). Meta's oversight board: a review and critical assessment. *Minds and Machines, 33*(2), 261-284.

World Economic Forum (2024). Metaverse Identity: Defining the Self in a Blended Reality. Insight Report March 2024. 1-48.

Wouters, J. et al. (2020). From an economic community to a union of values: The emergence of the EU's commitment to human rights. *The European Union and Human Rights*, 11-38.

Wu, H., Hui, P., & Zhou, P. (2023). Deepfake in the Metaverse: An Outlook Survey. *arXiv preprint arXiv:2306.07011*.

Yılmaz, H. K. E. (2024). Legal Issues Of The Metaverse: A Public International Law Perspective. *Law and Justice Review*, (27), 29-58.

Zurek, T., & Araszkiewicz, M. (2013). Modeling teleological interpretation. In *Proceedings of the fourteenth international conference on artificial intelligence and law*, 160-168.

Websites

Ahvenainen, J. (2022), 'Metaverses are coming, but who owns your avatar?' retrieved from: https://medium.com/prifina/Metaverses-are-coming-but-who-owns-your-avatar-61ae9750f9c2 on 18 April 2024.

Aitchison, M (2023), 'Aussie student's X-rated horror after innocently Googling her own name to discover someone had done the unthinkable - and her life will never be the same again' retrieved from: https://www.dailymail.co.uk/news/Article-11981501/Aussie-students-horror-Googling-life-never-again.html on 8 April 2024.

Amnesty International (2018), 'Toxic Twitter – the Silencing Effect' retrieved from: https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-Chapter-5-5/#:~:text=It%20states%2C,integrity%20of%20the%20information%20space%E2%80%A6 on 9 April 2024.

Belloni, P. (2024), 'The Proposal for a Directive on Gender-Based Violence'. Retrieved from: https://www.medialaws.eu/the-proposal-for-a-directive-on-gender-based-violence/ on 18 June 2024.

Compton, S. & Hamlyn, R. (2023), 'Opinion: The rise of deepfake pornography is devastating for women' retrieved from: https://edition.cnn.com/2023/10/29/opinions/deepfake-pornography-thriving-business-compton-hamlyn/index.html#:~:text=This%20practice%20is%20no%20longer,on%20victims%20can%20be%20devastating on 8 March 2024.

Contreras, B (2024), 'Tougher AI Policies Could Protect Taylor Swift—And Everyone Else—From Deepfakes. Retrieved from: https://www.scientificamerican.com/article/tougher-ai-policies-could-protect-taylor-swift-and-everyone-else-from-deepfakes/ 4 March 2024.

Council of Europe. (2024), 'Committee on Artificial Intelligence (CAI)'. Retrieved from: https://www.coe.int/en/web/artificial-intelligence/cai on 18 June 2024.

Council of Europe, 'The Budapest Convention (ETS No. 185) and its Protocols'. Retrieved from: https://www.coe.int/en/web/cybercrime/the-budapest-convention on 23 June 2024.

Council of Europe, 'Cybercrime Convention Committee'. Retrieved from: https://www.coe.int/en/web/cybercrime/tcy on 23 June 2024.

Council of Europe, 'Details of Treaty No.189'. Retrieved from: https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=189 on 23 June 2024.

Council of Europe, 'GREVIO'. Retrieved from: https://www.coe.int/en/web/istanbul-convention/grevio on 24 June 2024.

Council of Europe, 'Hate Crime and Hate Speech'. Retrieved from: https://rm.coe.int/thematic-factsheet-hate-crime-eng-docx/1680a96865 on 8 July 2024.

Council of Europe, 'International case law'. Retrieved from: https://www.coe.int/en/web/cyber violence/international-case-law#{%22212947575%22:[0]} on 1 July 2024.

Council of Europe, ‘Key facts about the Istanbul Convention’. Retrieved from: https://www.coe.int/en/web/istanbul-convention/key-facts on 2 July 2024.

Council of the EU, (2021), ‘What is illegal offline should be illegal online: Council agrees position on the Digital Services Act’. Retrieved from: https://www.consilium.europa.eu/en/press/press-releases/2021/11/25/what-is-illegal-offline-should-be-illegal-online-council-agrees-on-position-on-the-digital-services-act/#:~:text=The%20rules%20set%20out%20under,should%20also%20be%20illegal%20online on 28 June 2024.

Council of the EU (2023), ‘Combatting violence against women: Council adopts decision about EU’s accession to Istanbul Convention’ retrieved from: https://www.consilium.europa.eu/en/press/press-releases/2023/06/01/combatting-violence-against-women-council-adopts-decision-about-eu-s-accession-to-istanbul-convention/ on 29 April 2024).

Council of the EU (2024), ‘Press Release: European digital identity (eID): Council adopts legal framework on a secure and trustworthy digital wallet for all Europeans’. Retrieved from: https://www.consilium.europa.eu/en/press/press-releases/2024/03/26/european-digital-identity-eid-council-adopts-legal-framework-on-a-secure-and-trustworthy-digital-wallet-for-all-europeans/ on 10 June 2024.

Cubewealth (2023), ‘What is the Metaverse: Origins, Platforms, Future, Warnings?’ retrieved from: https://www.bankoncube.com/post/what-is-the-Metaverse on 18 March 2024.

El Atillah, I. (2023), ‘'Living a lifelong sentence': How AI is trapping women in a deepfake porn hell” retrieved from: https://www.euronews.com/next/2023/04/22/a-lifelong-sentence-the-women-trapped-in-a-deepfake-porn-hell on 8 April 2024.

EU Artificial Intelligence Act, ‘Timeline of Developments’ retrieved from: https://artificialintelligenceact.eu/developments/ on 28 June 2024.

EU Fundamental Rights Agency (FRA), 'EU Charter of Fundamental Rights'. Retrieved from: https://fra.europa.eu/en/eu-charter#:~:text=The%20Charter%20of%20Fundamental%20Rights%20of%20the%20European%20Union%20(CFREU,the%20scope%20of%20EU%20law. On 2 July 2024.

European Commission, 'Data Protection in the EU'. Retrieved from: https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en#:~:text=The%20General%20Data%20Protection%20Regulation%20(GDPR),-Regulation%20(EU)%202016&text=A%20single%20law%20will%20also,applies%20since%2025%20May%202018. On 6 July 2024.

European Commission & Organization for Security and Co-operation in Europe (2022), 'Report on novelties in the 2018 revision of the Audiovisual Media Services Directive and proposed interventions into the Serbian Law on Electronic Media and the Law on Advertising'. Retrieved from: https://www.osce.org/files/f/documents/a/1/539114.pdf on 12 July 2024.

European Commission (2023), 'European Citizens' Virtual Worlds Panel: A new phase of citizen engagement'. Retrieved from: https://citizens.ec.europa.eu/virtual-worlds-panel_en#:~:text=The%20expected%20outcome%20of%20the,an%20initiative%20on%20the%20topic. On 12 June 2024.

European Commission (2024), 'Additional obligations for Very Large Online Platforms kick in for Pornhub, Stripchat and XVideos under the DSA'. Retrieved from : https://digital-strategy.ec.europa.eu/en/news/additional-obligations-very-large-online-platforms-kick-pornhub-stripchat-and-xvideos-under-dsa#:~:text=19%20April%202024-,Additional%20obligations%20for%20Very%20Large%20Online%20Platforms%20kick%20in%20for,Digital%20Services%20Act%20(DSA). On 29 April 2024.

European Parliament (2024), 'Factsheet Personal Data Protection'. Retrieved from: https://www.europarl.europa.eu/ftu/pdf/en/FTU_4.2.8.pdf on 1 May 2024.

Facebook, 'Not Without My Consent: A guide to reporting and removing intimate images shared without your consent'. Retrieved from: https://about.fb.com/wp-content/uploads/2017/03/not-without-my-consent.pdf on 28 June 2024.

Faciaai, 'How did Paul Walker appear in Fast 7 after his death?'. Retrieved from: https://faciaai.medium.com/how-did-paul-walker-appear-in-fast-7-after-his-death-fda6acfea096 on 13 March 2024.

FBI (2023), 'Public Service Announcement: Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes; retrieved from: https://www.ic3.gov/Media/Y2023/PSA230605 on 8 April 2024.

Forbes, 'A Short History of the Metaverse'. Retrieved from: https://www.forbes.com/sites/bernardmarr/2022/03/21/a-short-history-of-the-Metaverse/ on 18 March 2024.

Gartner (2024), 'Gartner Predicts 25% of People Will Spend At Least One Hour Per Day in the Metaverse by 2026'. Retrieved from: https://www.gartner.com/en/newsroom/press-releases/2022-02-07-gartner-predicts-25-percent-of-people-will-spend-at-least-one-hour-per-day-in-the-metaverse-by-2026 on 9 July 2024.

Grady, P. (2023), 'EU proposals will fail to curb nonconsensual deepfake porn*'. Retrieved* from: *https://datainnovation.org/2023/01/eu-proposals-will-fail-to-curb-nonconsensual-deepfake-porn/#:~:text=Existing%20and%20proposed%20laws%20will,material%20without%20the%20individual's%20consent*. on 17 April 2024.

Guney, G. (2022), 'The Istanbul Convention: A Missed Opportunity in Mainstreaming Cyber violence against Women in Human Rights Law?' retrieved from: https://www.ejiltalk.org/the-istanbul-convention-a-missed-opportunity-in-mainstreaming-cyber violence-against-women-in-human-rights-law/ on 2 May 2024.

Home security heroes (2023), 'State of Deepfakes'. Retrieved from: https://www.homesecurityheroes.com/state-of-deepfakes/ on 13 March 2024.

Jee, C (2020), ‘An Indian politician is using deepfake technology to win new voters’. Retrieved from: https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/ on 8 April 2024.

Kitsara, I. (2024), ‘The Metaverse and its impact on human rights, the rule of law and democracy’. Abridged version of the report for the Council of Europe’. Retrieved from: https://rm.coe.int/the-Metaverse-impact-on-and-its-impact-on-human-rights-the-rule-of-law/1680ae6bce on 4 March 2024.

Krishnan, M. (2023), ‘Can India tackle deepfakes?’. Retrieved from: https://www.dw.com/en/can-india-tackle-deepfakes/a-67791106 on 9 April 2024.

Levy, S. (2022), ‘What’s Deepfake Bruce Willis Doing in My Metaverse?’. Retrieved from: https://www.wired.com/story/plaintext-bruce-willis-deepfake-Metaverse/ on 18 March 2024.

Mack, O. (2019), ‘Who Owns Your Digital Twin? Not You—and Here’s Why That’s a Massive Problem’. Retrieved from: https://www.newsweek.com/who-owns-your-digital-twin-not-you-heres-why-thats-massive-problem-opinion-1451991 on 5 July 2024 on 5 July 2024.

Macmahon, L. (2024), ‘Plans to use Facebook and Instagram posts to train AI criticised’ retrieved from: https://www.bbc.com/news/articles/cw99n3qjeyjo on 11 June 2024.

Marr, B. (2024), ‘The Metaverse And Its Dark Side: Confronting The Reality Of Virtual Rape’. Retrieved from: https://www.forbes.com/sites/bernardmarr/2024/01/16/the-Metaverse-and-its-dark-side-confronting-the-reality-of-virtual-rape/ on 18 March 2024.

Meta (2021), ‘Building the Metaverse responsibly’. Retrieved from: https://about.fb.com/news/2021/09/building-the-Metaverse-responsibly/ on 2 July 2024.

Meta (2024), ‘Code of Conduct for Virtual Experiences’. Retrieved from: https://www.meta.com/en-gb/help/quest/articles/accounts/privacy-information-and-settings/code-of-conduct-for-virtual-experiences/ on 2 July 2024.

Meta (2024), ‘Safety and privacy tools available for teens in Meta Horizon Worlds’. Retrieved from https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/safety-and-privacy-tools-teens-horizon-worlds/ on 2 July 2024.

Meta, ‘Transparency reports’. retrieved from: https://transparency.meta.com/reports/ on 3 July 2024.

Milmo, D. (2021), ‘Enter the Metaverse: the digital future Mark Zuckerberg is steering us toward’ retrieved from: https://www.theguardian.com/technology/2021/oct/28/facebook-mark-zuckerberg-meta-Metaverse on 4 March 2024.

Molloy, S. (2019), ‘Blackmailers-for-hire ae weaponising ‘deepfake’ revenge porn’ retrieved from https://nypost.com/2019/01/02/blackmailers-for-hire-are-weaponizing-deepfake-revenge-porn/ on 8 April 2024.

Mooij, A. & Tushuizen, J. (2024), ‘Regulating the Virtual World as a new State’. Retrieved from: https://europeanlawblog.eu/2024/06/28/regulating-the-virtual-world-as-a-new-state/ on 1 July 2024.

Mort, H. (2022), ‘The images are seared onto my retinas - I felt ashamed’: The poet who was a victim of deepfake porn. Retrieved from: https://www.telegraph.co.uk/women/life/images-seared-onto-retinas-felt-ashamed-poetvictim-deepfake/ on 8 April 2024.

Muncaster, P. (2023), ‘Deepfaking it: What to know about deepfake-driven sextortion schemes’ retrieved from: https://www.welivesecurity.com/2023/07/04/deepfaking-it-deepfake-driven-sextortion-schemes/ on 8 April 2024.

Nilesh, C. (2020), ‘We've Just Seen the First Use of Deepfakes in an Indian Election Campaign’. Retrieved from: https://www.vice.com/en/Article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp on 8 April 2024.

None of Your Business (NOYB) (2024), 'NOYB fordert 11 Behörden auf, Metas Missbrauch persönlicher Daten für KI zu stoppen'. Retrieved from: https://noyb.eu/de/noyb-urges-11-dpas-immediately-stop-metas-abuse-personal-data-ai on 2 July 2024.

Openletter.net (2024), 'Deepfakes'. Retrieved from: https://openletter.net/l/disrupting-deepfakes on 22 June 2024.

Oswego, 'Social Identity'. Retrieved from: https://ww1.oswego.edu/diversity/day-3-social-identity#:~:text=Personal%20Identity%20markers%20are%20often,something%20we%20are%20born%20into. On 1 July 2024.

Oversight Board, 'Mission statement'. Retrieved from: https://www.oversightboard.com/ on 1 July 2024.

Pellegrini, C. (2023), 'Conflict of Laws and the Metaverse'. Retrieved from: https://eapil.org/2023/06/13/conflict-of-laws-and-the-Metaverse/ on 4 March 2024.

Ramos, T. a.o. (2024), 'Navigating New Realities: The Impact of the revised eIDAS Regulation on the Metaverse and VLOPs' retrieved from: https://www.taylorwessing.com/en/insights-and-events/insights/2024/03/embracing-the-future-of-digital-identities on 10 June 2024.

Rodeck, D., & Curry, B. (2022), 'What is blockchain.' Retrieved from: https://www.forbes.com/advisor/in/investing/cryptocurrency/what-is-blockchain/ on 5 July 2024.

Rozynek, M. (2022), 'Me, myself and my avatar: Data ownership in virtual worlds'. Retrieved from: https://atelier.net/insights/me-myself-and-my-avatar-data-ownership-in-virtual on 12 July 2024.

RTL Nieuws, 'Werkstraf van 120 uur geëist voor deepfakevideo Welmoed Sijtsma' retrieved from: https://www.rtl.nl/nieuws/artikel/5414062/werkstraf-geeist-voor-deepfake-welmoed-sijtsma on 4 March 2024.

Safi, A., Atack, A. (2024), 'Revealed: the names linked to ClothOff, the deepfake pornography app' retrieved from: https://www.theguardian.com/technology/2024/feb/29/clothoff-deepfake-ai-pornography-app-names-linked-revealed on 13 March 2024.

Saner, E. (2024), 'Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'. Retrieved from: https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box+on+4+March+2024 on 4 March 2024.

Shoemaker, P. (2022), 'What Is Digital Identity in the Metaverse?' Retrieved from: https://www.identity.com/identity-in-the-Metaverse/#Identity_in_the_Metaverse on 11 June 2024.

Sidley (2024), 'EU Formally Adopts World's First AI Law'. Retrieved from: https://datamatters.sidley.com/2024/03/21/eu-formally-adopts-worlds-first-ai-law/#:~:text=On%20March%2013%2C%202024%2C%20the,in%20favor%20of%20the%20legislation on 29 April 2024.

Speed, R. (2024), 'Meta faces multiple complaints in Europe over plans to train AI on user data'. Retrieved from: https://www.theregister.com/2024/06/06/meta_ai_complaints/ on 2 July 2024.

Stanford Encyclopaedia of Philosophy. 'Personal identity'. Retrieved from: https://plato.stanford.edu/entries/identity-personal/ on 1 July 2024.

Tenbarge, K. 'Found through Google, bought with Visa and Mastercard: Inside the deepfake porn economy'. Retrieved from: https://www.nbcnews.com/tech/internet/deepfake-porn-ai-mr-deep-fake-economy-google-visa-mastercard-download-rcna75071 on 8 March 2024.

Trail, ‘EU AI Act: How risk is classified. Retrieved from: https://www.trail-ml.com/blog/eu-ai-act-how-risk-is-classified on 28 June 2024.

Tsalidis, A. (2024), ‘Disrupting the Deepfake Pipeline in Europe’. Retrieved from: https://futureoflife.org/ai-policy/disrupting-the-deepfake-pipeline-in-europe/ on 9 April 2024.

Vescent H, (2022). ‘The Metaverse: A Missed Opportunity for Data Ownership and Privacy?' Retrieved from: https://www.biometricupdate.com/202201/the-metaverse-a-missed-opportunity-for-data-ownership-and-privacy on 5 July 2024.

Wanjiru, M. (2023), ‘Beyond Borders: Understanding The Trends of Internet Jurisdiction’. Retrieved from: https://paradigmhq.org/beyond-borders-understanding-the-trends-of-internet-jurisdiction/ on 3 July 2024.

White, M. (2022). ‘Abuse and harassment on the blockchain’ retrieved from: https://blog.mollywhite.net/abuse-and-harassment-on-the-blockchain/ on 4 March 2024.

Wolford, B., ‘What is GDPR, the EU’s new data protection law?’ Retrieved from: https://gdpr.eu/what-is-gdpr/ on 29 April 2024.

Woollacott, E. (2022). ‘Rise of deepfakes: who can you trust in the Metaverse?’ retrieved from https://cybernews.com/security/rise-of-deepfakes/ on 4 March 2024.

XRToday (2021), ‘Unpacking Meta: Where did the Word Metaverse come from?’. Retrieved from https://www.xrtoday.com/virtual-reality/unpacking-meta-where-did-the-word-Metaverse-come-from/ on 18 March.

Yildrim, B., & Aydinli, C. (2019), ‘Deepfake: An Assessment From The Perspective Of Data Protection Rules’ Retrieved from: https://www.mondaq.com/turkey/privacy-protection/863064/deepfake-an-assessment-from-the-perspective-of-data-protection-rules on 5 July 2024.

Jurisprudence

European Court of Human Rights

Handyside v. the United Kingdom, No. 5493/72, 7 December 1976.

X and Y v. the Netherlands, No. 8978/80, 26 March 1985.

Observer and Guardian v. the United Kingdom, No. 13585/88, 26 November 1991.

Y.F. v. Turkey, No. 24209/94, 22 July 2003.

Ilaşcu and Others v. Moldova and Russia, No. 48787/99, 8 July 2004.

D.H. and Others v. the Czech Republic, No. 57325/00, 13 November 2007.

E.S. and Others v. Slovakia, No. 8227/04, 15 September 2009.

Gäfgen v. Germany, No. 22978/05, 3 June 2010.

M.S.S. v. Belgium and Greece, No. 30696/09, 21 January 2011.

Axel Springer AG v. Germany, No. 39954/08, 7 February 2012.

Putistin v. Ukraine, No. 16882/03, 21 November 2013.

Tamiz v. the United Kingdom, No. 3877/14, 19 September 2017.

Çakmak v. Turkey, No. 34872/09, 21 November 2017.

M.L. and W.W. v. Germany, Nos. 60798/10 and 65599/10, 28 June 2018.

Jishkariani v. Georgia, No. 18925/09, 20 September 2018.

Denisov v. Ukraine, No. 76639/11, 25 September 2018.

Milićević v. Montenegro, No. 27821/16, 6 November 2018.

Buturugă v. Romania, No. 56867/15, 10 February 2020.

Volodina v. Russia (no. 2), No. 40419/19, 14 September 2021.

M.L. v. Slovakia, No. 34159/1714, October 2021.

Hurbain v. Belgium, No. 57292/16, 4 July 2023.

Court of Justice of the European Union

C-131/12 *Google Spain SL, Google Inc. / Agencia Española de Protección de Datos (AEPD), Mario Costeja González* ECLI:EU:C:2014:317.

C-362/14, *Maximillian Schrems / Data Protection Commissioner* ECLI:EU:C:2015:650.

C-18/18 *Eva Glawischnig-Piesczek / Facebook* ECLI:EU:C:2019:458.

C-311/18, *Data Protection Commissioner / Facebook Ireland Ltd, Maximillian Schrems* ECLI:EU:C:2020:559.

Legal Instruments

European Union

European Union, *Consolidated version of the Treaty on European Union*, OJ C326/15, 26.10.2012.

European Union, *Consolidated version of the Treaty on the Functioning of the European Union*, OJ C326/47, 26.10.2012.

European Parliament, the Council and the Commission, *Charter of Fundamental Rights of the European Union*, OJ C202/02, 7.6.2016.

European Parliament and Council, *Regulation 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation*), OJ L 119, 4.5.2016.

European Parliament and Council, *Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive*, OJ L 303, 28.11.2018.

European Parliament and of the Council, *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act*), OJ L 277, 27.10.2022.

European Parliament, *European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))*, OJ L, 2024/1689, 13.03.2024.

European Parliament and Council, *Regulation (EU) 2024/1183 of the European Parliament and of the Council of 11 April 2024 amending Regulation (EU) No 910/2014 as regards establishing the European Digital Identity Framework*, OJ L, 2024/1183, 30.4.2024.

European Parliament and Council, *Directive (EU) 2024/1385 of the European Parliament and of the Council of 14 May 2024 on combating violence against women and domestic violence*, OJ L, 2024/1385, 24.5.2024.

European Parliament and of the Council, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance*, OJ L, 2024/1689, 12.7.2024.

Council of Europe

Council of Europe (1950). European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14. ETS No. 161.

Council of Europe (2001). Convention on Cybercrime (Budapest Convention). ETS No. 185.

Council of Europe (2003). Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems. ETS No. 189.

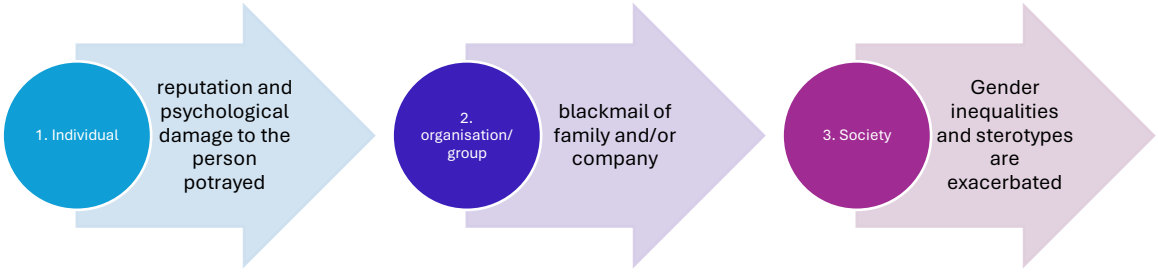
Council of Europe Convention on preventing and combating violence against women and domestic violence (Istanbul Convention) (2011). CETS No. 210.

Council of Europe (2022). Second Additional Protocol to the Convention on Cybercrime on enhanced co-operation and disclosure of electronic evidence. CETS No. 224.

Council of Europe (2024). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. CETS No. 224.

Annex

Figure 1:



Graph inspired by: Huijstee, M. V., et al. (2021). Tackling deepfakes in European policy. Think Tank European Parliament. *The European Parliament*. p. V.

Figure 2:

Mr. Dr. Bart W. Schermer & Joas van Ham MSc (2021). 'Regulering van immersieve technologieën Wetenschappelijk Onderzoek- en Documentatiecentrum' retrieved from: https://open.overheid.nl/documenten/ronl-d81ef4594c8bcca4b8e04ec659d0b6930f2cad67/pdf

Figure 3:

Ng, D. T. K. (2022). What is the metaverse? Definitions, technologies and the community of inquiry. *Australasian Journal of Educational Technology*, *38*(4), 190-205, p. 200.

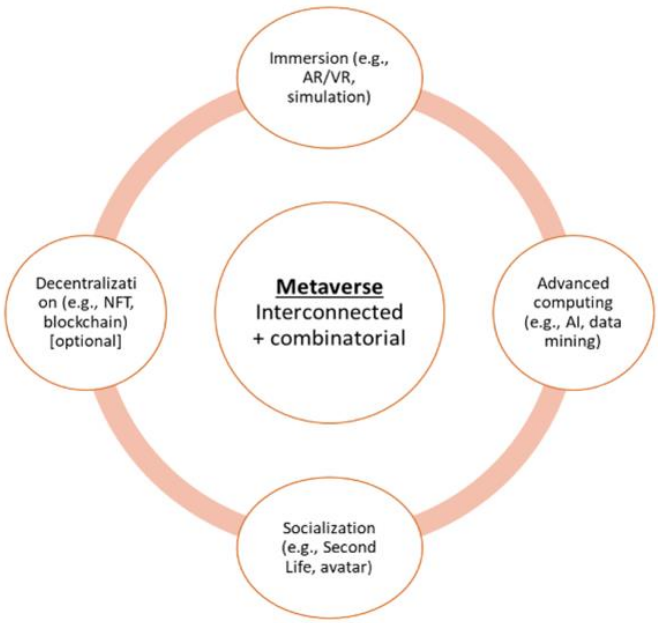
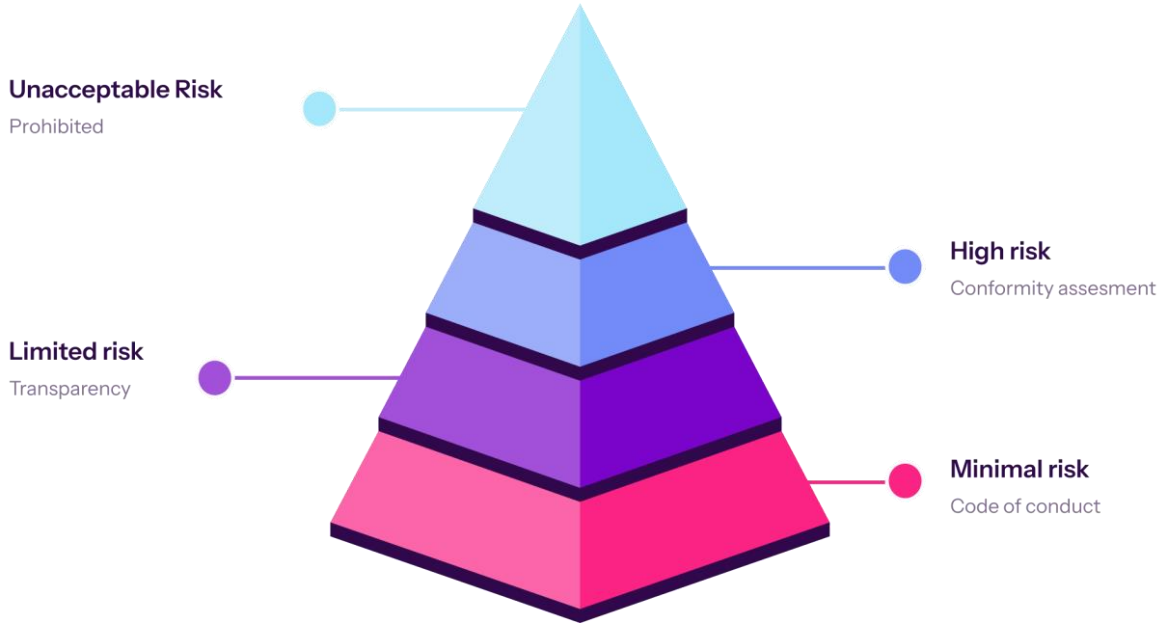


Figure 4:



Risk-based approach in the AIA. Image pyramid retrieved from: https://www.credo.ai/eu-ai-act.