Harmful deepfakes and the GDPR

how data protection legislation should be used to address issues revolved around the harmful use of deepfake technology

Master's Thesis Law and Technology
Tilburg Law School – Institute for Law, Technology and Society
Tilburg University (Spring intake 2021)
17.200 words

M.J. van der Helm (u513155)
10 December 2021

TFG: A Brave New World
Supervisor: M. Paun
Second reader: M.E. Noorman

Table of contents

Chapter 1: Introduction

1.1 Problem statement

Rapid developments in the field of Artificial Intelligence (hereinafter "**AI**") have made it possible to create fake audio-visual images of people, which are so realistic that they are indistinguishable from real ones.[1] These "deepfakes" can be used in movies to bring deceased actors back to life or to make them appear younger. It should not come as a surprise that this technology can also be used for harmful purposes. A shocking report from September 2019 found that approximately 96% of online circulating deepfake videos contained non-consensual pornographic material.[2] This report was created by Deeptrace,[3] which develops software for detecting fake online content.

According to that same report, the number of detected deepfakes increased from 7,964 in December 2018 to 14,678 by July 2019.[4] In the meantime, these numbers have increased to 24,263 by December 2019, 49,081 by June 2020 and 85,047 by December 2020.[5] The amount of detected deepfakes thus doubles roughly every six months which also means the number of videos containing non-consensual pornographic material is on the rise. Studies show that emotional harms of sexual privacy intrusions can be severe and lasting.[6] Victims have shown difficulty eating, working, and concentrating and experience anxiety and depression. Some even contemplate suicide.[7]

Apart from non-consensual pornography, deepfake technology can be used for many other harmful activities, ranging from stealing someone's identity for (financial) benefit[8] to manipulating elections.[9] Deepfakes created for these harmful purposes are problematic for many reasons. They cause reputational harms to those depicted.[10] This could result in victims losing their jobs and having trouble finding new ones because employers might worry that

[1] M.B. Kugler and C.L. Pace, 'Deepfake Privacy: Attitudes and Regulation' (2021) 116 Northwestern University Law Review, p. 10, available at: <https://ssrn.com/abstract=3781968>.
[2] H. Adjer. G. Patrini, F. Cavalli and L. Cullen, 'The State of Deepfakes: Landscape, Threats and Impact' (2019), p. 1, <https://regmedia.co.uk/2019/10/08/deepfake_report.pdf> last accessed 11 November 2021.
[3] Deeptrace is now called Sensity.
[4] Ibid.
[5] F. Cavalli, 'How to detect a deepfake online' (*Sensity*, 8 February 2021) <https://sensity.ai/how-to-detect-a-deepfake/> last accessed 11 November 2021.
[6] D.K. Citron, 'Sexual Privacy' (2019) 128 Yale Law Journal 1870, p. 1926.
[7] Ibid.
[8] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (2019) 107 California Law Review 1753, p. 1772.
[9] Ibid, p. 1778.
[10] Ibid, p. 1774.

their employee's reputation might reflect badly on them.[11] Moreover, victims may suffer all sorts of psychological damages, as already explained. Harmful deepfakes are also privacy intrusive as they are often created without consent.

Since deepfakes are relatively new, many countries have not yet regulated them in targeted legislation. Targeted legislation could be an effective way to deal with the problems deepfakes may cause because targeted legislation regulates the use of deepfake technology itself. However – at the same time – we can ask ourselves whether existing legislation could (rather) be used to deal with these issues. Within the European Union (hereinafter: "**EU**"), the General Data Protection Regulation[12] (hereinafter "**GDPR**") could serve as a starting point for regulating deepfakes. This is because personal data is almost always processed when deepfakes are created[13] and the GDPR has a wide reach and applicability,[14] as will be demonstrated throughout this thesis. Therefore, various rights and remedies can possibly be used for protection.[15] I believe that if personal data has the potential to impact individuals, some form of legal protection should be triggered.[16] The GDPR is supposed to protect individuals' personal data,[17] which makes it a good starting point.

The goal of my thesis is two-fold. On the one hand, it is meant to explore how the GDPR should be used to regulate harmful deepfakes by looking at the different provisions of the GDPR and assessing which remedies should be used and to what extent they offer protection. On the other hand, my thesis is meant to evaluate whether the GDPR should be used in the first place to regulate harmful deepfakes and if so, what its role should be in light of other regulatory solutions such as criminal law and targeted legislation. I will clarify my choice for criminal law and targeted legislation in section 1.6.

[11] D.K. Citron, *Hate Crimes in Cyberspace* (2014 Harvard University Press), p. 7-8.

[12] Regulation (EU) 2016/679 of the European Parliament and 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 [2016] OJ L119/1 (hereinafter: "**GDPR**").

[13] European Parliament, 'Tackling deepfakes in European policy' (2021), p. 64 <https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf> last accessed 11 November 2021.

[14] See the wording of Article 3(1) GDPR.

[15] European Parliament, 'Tackling deepfakes in European policy' (2021), p. 39 <https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf> last accessed 11 November 2021. See also 'Communications from the Commission to the European Parliament and the Council – data protection as a pillar of citizen's empowerment and the EU's approach to the digital transition- two year of application of the General Data Protection Regulation' COM(2020) 364final, p. 10, available at <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0264&from=EN> where it is noted that the GDPR has been designed in a technology neutral way, applying to new technologies as they develop. Therefore, the GDPR can be said to be designed to cover deepfakes, as a new technology.

[16] Argument derived from N. Purtova 'The law of everything. Broad concept of personal data and future of EU data protection law' (2018) Vol. 10 Law, Innovation and Technology 40, p. 74.

[17] Article 1 GDPR

As I will indicate in the next section, the topic of my thesis is quite unexplored. Therefore, the conclusion of my thesis could be valuable to an overall discussion revolving around the question how deepfakes can be regulated and particularly, what the role of the GDPR should be.

1.2 Existing literature

Since the term "deepfake" first emerged in 2017, [18] the yearly appearing number of scholarly papers addressing them has increased significantly.[19] There seems to be consensus on how deepfakes can be harmful. Robert Chesney and Danielle Citron explain that the harmful use of deepfakes can generally be divided into two categories.[20] Firstly, deepfakes can be used to cause harm to individuals or organizations by spreading fake content for purposes of exploitation or sabotage. Examples of deepfakes that could cause such dignitary harms[21] include non-consensual pornography, financial fraud or hate speech.[22] Secondly, deepfakes can be used to cause harm to society.[23] Examples include fake videos that feature public officials displaying racism, taking bribes, or engaging in adultery. In these cases, damage to society could extend to, among other things, distortion of democratic discourse on important policy questions and trust erosion in public institutions.[24]

Different solutions to deal with these issues have been discussed in the literature. Kelsey Farish, Rebecca Delfino and Augustine Eigbedion argue that certain legislative instruments could be used to regulate the use of deepfake technology. Examples include regulation

[18] See for example: B. Goggin, 'From porn to 'Game of Thrones': How deepfakes and realistic-looking fake videos hit big' (2019) <https://www.businessinsider.com/deepfakes-explained-the-rise-of-fake-realistic-videos-online-2019-6?IR=T> last accessed 11 November 2021.

[19] T.T. Nguyen, C.M. Nguyen, D.T Nguyen, D.T. Nguyen, S. Nahavandi, 'Deep Learning for Deepfakes Creation and Detection: A Survey' (2020), p. 2 <https://arxiv.org/pdf/1909.11573v2.pdf> last accessed 11 November 2021.

[20] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 California Law Review 1753.

[21] M.B. Kugler and C.L. Pace, 'Deepfake Privacy: Attitudes and Regulation' (2021) 116 Northwestern University Law Review, p. 10, available at: <https://ssrn.com/abstract=3781968>.

[22] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 California Law Review 1753, p. 1772.

[23] M.B. Kugler and C.L. Pace, 'Deepfake Privacy: Attitudes and Regulation' (2021) 116 Northwestern University Law Review, p. 11, available at: <https://ssrn.com/abstract=3781968>.

[24] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 California Law Review 1753, p. 1776-1777.

through data protection law[25], privacy law[26], criminal law[27], competition law[28] and copyright law.[29] Mika Westerlund and Oscar Schwartz mention soft law mechanisms like training,[30] corporate policies and voluntary actions by social media firms[31] as tools to regulate deepfakes. Besides that, the European Commission has proposed legislation that – among other things – shall impose "minimum transparency obligations" for users[32] of AI systems, to disclose that "content has been artificially generated or manipulated"[33] when using such AI systems to create deepfakes. Finally, Abbas Yazdinejad, Reza Parizi, Gautam Srivastava, and Ali Dehghatanha have explored to what extent Blockchain technology[34] could play a role in validating the authenticity of audio-visual images and filtering out deepfakes.

However, all but one of the cited papers leave the role of the GDPR undiscussed.[35] Sources that do explore the GDPR in relation to deepfakes are limited to weblogs[36] and a master thesis by Daphne Stevens.[37] These sources offer good overviews on certain remedies for victims of harmful deepfakes but lack extensive reviews on the role of the GDPR in light of other regulatory solutions. Besides that, the weblogs are not of academic nature and

[25] A. Eigbedion, 'Deepfakes: Legal & Regulatory Considerations in Nigeria' (2020), p 7-9, available at: <https://ssrn.com/abstract=3670644>.

[26] K. Farish, 'Do Deepfakes Pose a Golden Opportunity? Considering Whether English Law Should Adopt California's Publicity Right in the Age of Deepfake (2020) Vol. 15, No. 1 Journal of Intellectual Property Law & Practice 40, p. 44-46.

[27] R. Delfino, 'Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act' (2019), 88 Fordham Law Review 887, p. 926-928.

[28] A. Eigbedion, 'Deepfakes: Legal & Regulatory Considerations in Nigeria' (2020), p 11-12, available at: <https://ssrn.com/abstract=3670644>.

[29] E. Meskys, J. Kalpokiene, P. Jurcys, A. Liaudanskas, 'Regulating Deep-Fakes: Legal and Ethical Considerations' (2019) Vol. 15(1) Journal of Intellectual Property Law & Practice 24, p. 27.

[30] M. Westerlund, 'The Emergence of Deepfake Technology: A Review' (2019) Vol. 9(11) Technology Innovation Management Review 40.

[31] O. Schwartz, 'Deepfakes aren't a tech problem. They're a power problem' (2019) <https://www.theguardian.com/commentisfree/2019/jun/24/deepfakes-facebook-silicon-valley-responsibility> last accessed 11 November 2021.

[32] "Users" are defined as "any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity" according to article 3(4) of the Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Proposal for Artificial Intelligence Act) and amending certain Union legislative acts [2021] COM(2021) 206final.

[33] Article 52(3) Proposal for Artificial Intelligence Act.

[34] A. Yazdinejad, R. M. Parizi, G. Srivastava and A. Dehghantanha, 'Making Sense of Blockchain for AI Deepfakes Technology' (2020) GC Wkshps 2020 1.

[35] Of the cited papers, only the one written by K. Farish mentions GDPR. However, this is limited to a brief mentioning of two challenges for asserting one's GDPR rights regarding deepfakes.

[36] See for instance. B.C. Yildirim and C.D. Aydinli, 'Turkey: Deepfake: An Assessment From The Perspective Of Data Protection Rules' (*Mondaq*, 13 November 2019) <https://www.mondaq.com/turkey/privacy-protection/863064/deepfake-an-assessment-from-the-perspective-of-data-protection-rules> last accessed 11 November 2021 and M. Hallé, 'Deep fakes: are there remedies for victims under the GDPR?' (*International Bar Association*, 29 November 2018) <https://www.ibanet.org/Article/NewDetail.aspx?ArticleUid=3ED0EDC2-92E2-477B-8321-1290AFE00ACC> last accessed 15 April 2021.

[37] D. Stevens, 'Regulating Deepfake Technology' (2020), available at <https://arno.uvt.nl/show.cgi?fid=152071>.

Stevens' analysis of the GDPR is limited to a narrow analysis of the right to be forgotten laid down in article 17(1) GDPR. Consequently, there is no academic paper yet out there that conducts a substantive analysis of which remedies of the GPDR should be used to regulate harmful deepfakes and to what extent they offer protection.

Moreover, there are many authors that have written about the (role of the) GDPR, but none explore how it should be used to regulate harmful deepfakes, considering other regulatory solutions. Nadezhda Purtova – for example – does reflect on the general role of the GDPR in an age where almost everything entails personal data[38] and Tal Zarksy does not welcome its broad applicability and reach[39] but these authors do not discuss how the GDPR should be used to regulate harmful deepfakes specifically. Although Stevens does explore how the GDPR can regulate harmful deepfake pornography, she does not discuss how the GDPR should relate to other regulatory solutions, although she does mention them. The reason for this is that Stevens' focus is primarily on exploring which legal instruments can be used to regulate harmful deepfake pornography, without assessing how they should relate to each other.

The gap in the literature therefore is an extensive analysis of how the GDPR should be used to regulate the harmful use of deepfake technology. By answering the main research question and sub-questions that I have formulated in the next sections, my thesis will serve as a starting point to fill this gap. Writing about this topic is important to an overall discussion on how deepfakes can and should be regulated.

1.3 Main research question

I have identified that deepfakes can be used to cause harm to individuals. The objective of my thesis is to explore what the role of the GDPR should be in dealing with these issues. To achieve this objective, the main question that my thesis will revolve around answering is:

> *"How should the GDPR be used to offer protection against the harmful use of deepfake technology?"*

[38] See for instance: N. Purtova 'The law of everything. Broad concept of personal data and future of EU data protection law' (2018) Vol. 10 Law, Innovation and Technology 40.

[39] T. Zarsky, 'Incompatible: The GDPR in the Age of Big Data' (2017) Vol. 47, No. 4(2) Seton Hall Law Review 995.

1.4 Sub-questions

To answer the main research question, the following sub-questions must first be answered:

1. *What is deepfake technology and how can it be used to harm the interests of individuals?*
2. *To what extent do the main mechanisms of the GDPR protect against the harmful use of deepfake technology?*
3. *How could other regulatory frameworks such as criminal law and targeted legislation be used to address shortcomings of the GDPR when regulating the harmful use of deepfake technology?*

1.5 Methodology, methods, and structure

The objective of my research question is evaluative, aiming to establish how the GDPR should be used to regulate the harmful use of deepfake technology.

To answer my main research question, I will first explain what deepfake technology entails and how it can be used to harm the interests of individuals. Chapter 2 will thus be of explanatory nature. I will be going through secondary academic sources that are descriptive. This strategy is important for answering the first sub-question. In this chapter, I will explain how deepfakes are created, what they are used for and how they can cause harm to individuals. It is meant to show why it is important that deepfakes are regulated by giving insights on what problems they can cause.

Chapter 3 will contain a doctrinal legal analysis of the GDPR, which means I will conduct a descriptive and detailed analysis of how the GDPR applies to deepfakes. In doing so, I will be able to assess which legal remedies offer the most protection to individuals. The method I will use to achieve this will mainly be a "black letter law approach" by looking at the different recitals and provisions of the GDPR and using case law and literature to explain the most relevant provisions for my thesis. This chapter will demonstrate to who victims can turn to exercise their rights under the GDPR. It will also highlight shortcomings of using the GDPR as a regulatory solution. In this way, I will answer the second sub-question.

I will be using a similar approach regarding chapter 4 to explore the role of the GDPR in light of other regulatory solutions and to assess how criminal law and targeted legislation could address the GDPR's shortcomings. My main sources will however be of secondary academic nature. I will start by exploring whether the GDPR should be used in the first place to regulate harmful deepfakes and if so, how it should relate to other regulatory frameworks.

In this way I will be able to demonstrate whether the GDPR's shortcoming should be addressed by applying other regulatory solutions alternatively or rather complementary. I will then look at criminal law and assess whether it offers solutions to regulatory gaps, I have identified in the GDPR's main mechanisms. I will also explore to what extent it is desirable that targeted legislation is adopted to regulate the harmful use of deepfake technology and what the advantages and disadvantages of this are. By taking this approach, I will be able to answer my third research question and shed light on how the GDPR should be used to regulate the harmful use of deepfake technology, in light of other regulatory solutions.

Chapter 5 will be used to answer the main research question and will contain the conclusion of this thesis.

1.6 Limitations in scope

I have narrowed down the scope of my research to the harmful use of deepfakes regarding individuals and thus will not be focussing on societal harms or harms to organizations. I deem this narrowing down appropriate because the GDPR sets rules for the protection of natural persons.[40] That does however not detract from the fact that the GDPR could also indirectly protect organisations and society. The term "harmful deepfake" encompasses both deepfakes that cause intentional and unintentional harms. The focus will however primarily be laid on deepfakes that have been created to intentionally harm individuals. This limitation in scope will allow me to focus primarily on the rights of victims rather than creators of deepfakes that have good intentions and want to search for guidance on GDPR compliance.

An important limitation regarding chapter 3 is my focus on social media platforms and search engines as actors before whom victims of harmful deepfakes can exercise their data subject rights. I define social media platform as a website or application which allows users to interact with each other.[41] Examples include YouTube and Facebook.[42] Search engines can be used to look for information online. An example is Google. I will not discuss data protection issues regarding the publication of deepfakes on the dark web or other non-transparent networks. Moreover, I will not be going into the possibility for a data subject to mandate non-

[40] Article 1 GDPR.

[41] Based how Caleb T. Carr and Rebbeca A. Hayes, define social media as "Internet-based channels that allow users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others" in C.T. Carr & R.A. Hayes, 'Social Media: Defining, Developing and Divining' 23(1) Atlantic Journal of Communication 46, p. 50.

[42] Ibid, p. 53.

profit-entities, to lodge complaints, file judicial remedies or claim damages on behalf of the data subject.[43]

Another limitation regards my choice in chapter 4 to explore criminal law and targeted legislation to deal with regulatory gaps of the GDPR. I will not discuss copyright law,[44] competition law, regular tort law or any other legal instruments. I am aware that these frameworks could also be used to regulate harmful deepfakes, but I cannot discuss them all, mainly because of limitations in time and space. I have chosen to look at criminal law and targeted legislation because I believe these two solutions are most effective in dealing with the GDPR's regulatory gaps. Concretely, my choice for criminal law is based on the general principle that it directly targets and prohibits harmful behaviour. Criminal law thus is constructed in such a way, that it targets harmful behaviour revolved around deepfake technology, meaning that it could be well-equipped to address harmful deepfakes. Therefore, it is an interesting alternative solution for regulating the harmful use of deepfake technology. As I will note in section 4.4, current regulatory frameworks are ill-equipped to effectively deal with harmful deepfakes. Therefore, I deem it important to also assess whether targeted legislation could be used to deal with regulatory gaps of the GDPR. I also believe it is interesting to look at the regulation of harmful deepfakes from different perspectives: the GDPR can be regarded technology-neutral,[45] criminal law could be seen as technology-independent,[46] and targeted legislation is rather technology-dependent, since it regulates a certain technology.

[43] Article 80 GDPR.

[44] This area of law has many limitations when used to regulate the harmful use of deepfake technology. For instance, it provides remedies for the copyright owner and not the victim. See for example E. Meskys, J. Kalpokiene, P. Jurcys, A. Liaudanskas, 'Regulating Deep-Fakes: Legal and Ethical Considerations' (2019) Vol. 15(1) Journal of Intellectual Property Law & Practice 24, pp. 27-28.

[45] Meaning it was designed to protect personal data, regardless of technology. See: European Commission, 'The GDPR: new opportunities, new obligations' (2018), p. 5 <https://ec.europa.eu/info/sites/default/files/data-protection-factsheet-sme-obligations_en.pdf> last accessed 11 November 2021. See also: wording of recital 15 GDPR.

[46] E.J Koops, M. Lips, C. Prins & M. Schellekens, 'Should ICT Regulation Be Technology neutral?' (2006) Vol. 9 IT & Law Series, p. 5 available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=918746>.

Chapter 2: Harmful use of deepfake technology

2.1 Introduction

This first substantive chapter will make clear what deepfake technology entails and how it can be used to bring harm to the interests of individuals. In this way, it is meant to illustrate what the problems are revolved around the use of the technology for harmful purposes and why victims need to be protected.

Section 2.2 will give a brief historic background on methods of manipulating multimedia content and will describe where the term "deepfake" comes from. Section 2.3 will explain the technology used to create deepfakes. Section 2.4 will discuss individual and societal benefits of deepfake technology. Section 2.5 will deal with different examples of how deepfake technology can be used to harm the interests of individuals. Section 2.6 will be used to answer the first sub-question.

2.2 Content manipulation techniques

One of the earliest examples of manipulated multimedia content is an 1860 portrait of a southern US politician named John Calhoun, who's head was replaced with the head of US president Abraham Lincoln.[47] This kind of manipulation is usually accomplished by adding, removing, and replicating objects between or within images. Then, post-processing steps like colour adjustments and scaling or rotating are applied to make the manipulated image look more authentic.[48]

Apart from these traditional methods of manipulating audio-visual content, advancements in computer technology have made it possible to manipulate digital content in very convincing ways.[49] One simple technique is to edit video footage by slowing it down, speeding it up, cutting parts out, or spicing parts together.[50] The result is that the original context of the video can easily be changed, which can have negative consequences for those featured. An example of this technique being used is a video from May 2019, where a seemingly drunk Nancy Poleski talking about Donald Trump appeared online. The "drunk

[47] F.Y. Shih and Y. Yuan 'A Comparison Study on Cover-Cover Image Forgery Detection' (2010) The Open Artificial Intelligence Journal 49, p. 49.

[48] L. Verdoliva, 'Media Forensics and Deepfakes: an overview' (2020), p. 2 <https://arxiv.org/pdf/2001.06564.pdf> last accessed 11 November 2021.

[49] M. Masood, M. Nawaz, K.M. Malik, A. Javed, A. Irtaza, 'Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward' (2021), para. 3 <https://arxiv.org/ftp/arxiv/papers/2103/2103.00484.pdf>.

[50] Ibid.

Nancy Poleski" clip caused public outcry, but it was quickly discovered that it was created by slowing down the original video to a speed of roughly 75%.[51] In the Netherlands, video-editing techniques have been used prior to past elections to mislead voters. For instance, videos of debates between different parties have been manipulated by cutting and pasting parts to make answers to questions seem foolish or false.[52]

Manipulation of existing video footage, where individuals are realistically depicted saying things they have never said started in 1997, when a paper of the "Video Rewrite Program" was published.[53] It contained the findings of a study of the very first software that could be used to reanimate facial movements in existing videos automatically to different audio tracks. The purpose of this program was to develop a technique to synthesize faces automatically and convincingly with audio tracks and was originally intended to be used in movies and thus not to cause (intentional) harm.[54] The program achieved convincing results and can be considered a landmark project, which ultimately led to the development of deepfake technology.

The first deepfake however, did not appear until 2017, when a Reddit user named "deepfakes" posted videos online in which the faces of famous actresses were swapped onto porn videos.[55] The term "deepfake" stems from a combination of "deep learning" and "fake." Today, anyone can create a deepfake video within seconds, using easily accessible online deepfake applications.

2.3 The technology behind deepfakes

The underlying mechanisms for creating deepfakes consist of machine learning algorithms,[56] which are used to insert (audio-visual) face images[57] of "target persons" into

[51] B. Paris and J. Donovan, 'Deepfakes and Cheap Fakes' (2019) p. 30 <https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal.pdf> last accessed 11 November 2021.

[52] 'Online filmpjes zijn het nieuwe politieke wapen, maar 'het effect is beperkt'' *rtlnieuws* (Hilversum, 25 september 2020) <https://www.rtlnieuws.nl/nieuws/politiek/artikel/5186252/verkiezingen-politiek-campagne-thierry-baudet-pieter-omtzigt-fvd> last accessed 11 November 2021.

[53] C. Bregler, M. Covell, M. Slaney, 'Video Rewrite: Driving Visual Speech with Audio' (proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, August 1997) <https://dl.acm.org/doi/pdf/10.1145/258734.258880> p. 353-360.

[54] Ibid, p. 353.

[55] See for instance: B. Goggin, 'From porn to 'Game of Thrones': How deepfakes and realistic-looking fake videos hit big' (2019) <https://www.businessinsider.com/deepfakes-explained-the-rise-of-fake-realistic-videos-online-2019-6?IR=T> last accessed 11 November 2021.

[56] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (2019) 107 California Law Review 1753, p. 1759.

[57] E. Meskys, J. Kalpokiene, P. Jurcys, A. Liaudanskas, 'Regulating Deep-Fakes: Legal and Ethical Considerations' (2019) Vol. 15(1) Journal of Intellectual Property Law & Practice 24, p. 25. Audio or video can also be inserted. See for example: C.Q. Choi, 'AI Creates Fake Obama > Videos of Barack Obama made from existing audio, video of him' (2017) <https://spectrum.ieee.org/ai-creates-fake-obama> last accessed 11 November 2021.

video image(s) of "source persons." The technology generally involves the use of neural networks, which are computing systems that are inspired by biological neural networks in human brains. When multiple neural network layers are used, this can be referred to as "deep learning."[58] During deep learning, the connections in the neural network are strengthened or weakened, which makes the system better at making accurate predictions from input data.[59] It is through deep learning that neural networks are able to categorize images, audio or video to generate realistic fake audio-visual content.[60] Deep learning techniques is responsible for the best-performing AI systems we have today.[61]

Like the human brain, predictions become more accurate through experience.[62] In deep learning algorithms, experience is gained by adding more input data. These algorithms will be able to create increasingly accurate models, the more data is used. For this reason, public figures like celebrities and politicians were the initial targets of deepfakes, since they have many videos and images available online.[63] However, less data will be required to create a realistic deepfake in the future, as the technology is becoming more efficient.[64] For instance, the phone app "WOMOBO" requires users to insert just one photo to create a deepfake of someone singing a popular song.[65] The results are not very convincing, but one can imagine how easy it will be to create a more realistic deepfake of someone with only little input data, as deepfake technology improves. The result is that anyone can become a target for a deepfake, especially considering how easy it is to get a hold of someone's photos through social media.

[58] L. Hardestry, 'Explained: Neural networks – Ballyhooed artificial-intelligence technique known as "deep learning" revives 70-year-old idea (2017), <https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414> last accessed 11 November 2021.

[59] J.B. Heaton, N.G. Polson, J.H. Witte, 'Deep Learning for Finance: Deep Portfolios' (2016) Vol. 33(1) Applied Stochastic Models in Business and Industry 3, p. 4.

[60] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (2019) 107 California Law Review 1753, p. 1760.

[61] L. Hardestry, 'Explained: Neural networks – Ballyhooed artificial-intelligence technique known as "deep learning" revives 70-year-old idea (2017), <https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414> last accessed 11 November 2021.

[62] Ibid, p. 1759.

[63] T.T. Nguyen, C.M. Nguyen, D.T Nguyen, D.T. Nguyen, S. Nahavandi, 'Deep Learning for Deepfakes Creation and Detection: A Survey' (2020), p. 1 <https://arxiv.org/pdf/1909.11573v2.pdf> last accessed 11 November 2021.

[64] E. Meskys, J. Kalpokiene, P. Jurcys, A. Liaudanskas, 'Regulating Deep-Fakes: Legal and Ethical Considerations' (2019) Vol. 15(1) Journal of Intellectual Property Law & Practice 24, p. 25.

[65] Another example is the website thispersondoesnotexist.com, which was trained with tens of thousands of online photos (see This person does not exist: AI generates fake faces on website' (2019) <https://www.ctvnews.ca/sci-tech/this-person-does-not-exist-ai-generates-fake-faces-on-website-1.4299515> last accessed 11 November 2021) as input data to create realistic images of human faces (output data), which are indistinguishable from photos of real human faces. Try <whichfaceisreal.com> to see how difficult it is to tell images of real human faces apart from AI-generated images.

Using deep learning algorithms in the way described above thus makes it possible to create very realistic fake audio-visual content. However, to create a deepfake that is undistinguishable from a real video, more sophisticated methods should be used. Examples of such methods include "autoencoders" and "generative adversarial networks" (hereinafter: "**GANs**").[66] Both approaches have their own strengths and weaknesses. Autoencoders are easily trained but generally produce blurry results. GANs produce sharp images, but only in small resolutions.[67]

The first application for deepfake creation was "FakeApp", which was developed using an autoencoder-decoder pairing structure.[68] In this method, the hidden features of face images are extracted by the autoencoder and reconstructed by a decoder. To swap someone's face, two encoder-decoder pairs are used, of which the encoders are part of the same network.[69] In this way, the encoder will be able to learn differences and similarities of a source image and a target image. This is relatively easily done, since faces have similar features such as nose, eyes, and mouth positions.[70]

Creating deepfakes through GANs is a more popular method[71] because the results are more realistic.[72] A GAN usually consists of two neural networks. One network is known as the "generator," which generates a data sample (for instance, fake image of a human) that can pass for real data. The other network, the discriminator, tries to assess which data is original and which data has been generated by the generator. The GAN approach was invented by Ian Goodfellow, who uses a team of counterfeits that produces fake currency as an analogue to describe the generatic model (generator) and the police trying to detect the fake money as an analogue to the discriminative model (discriminator). Both teams are driven to improve their

[66] T.T. Nguyen, C.M. Nguyen, D.T Nguyen, D.T. Nguyen, S. Nahavandi, 'Deep Learning for Deepfakes Creation and Detection: A Survey' (2020), p. 1 <https://arxiv.org/pdf/1909.11573v2.pdf> last accessed 11 November 2021. For an overview of different types of methods see M. Masood, M. Nawaz, K.M. Malik, A. Javed, A. Irtaza, 'Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward' (2021) <https://arxiv.org/ftp/arxiv/papers/2103/2103.00484.pdf>.

[67] T. Karras, T. Aila, S. Lainen, J. Lehtinen, 'Progressive Growing of GANs for Improved Quality, Stability, and Variation' (International Conference on Learning Representations, Vancouver, April-May 2018), p. 1.

[68] T.T. Nguyen, C.M. Nguyen, D.T Nguyen, D.T. Nguyen, S. Nahavandi, 'Deep Learning for Deepfakes Creation and Detection: A Survey' (2020), p. 1 <https://arxiv.org/pdf/1909.11573v2.pdf> last accessed 11 November 2021.

[69] Ibid, p. 2

[70] Ibid.

[71] H. Adjer. G. Patrini, F. Cavalli and L. Cullen, 'The State of Deepfakes: Landscape, Threats and Impact' (2019), p. 3, <https://regmedia.co.uk/2019/10/08/deepfake_report.pdf> last accessed 11 November 2021.

[72] T. Karras, T. Aila, S. Lainen, J. Lehtinen, 'Progressive Growing of GANs for Improved Quality, Stability, and Variation' (International Conference on Learning Representations, Vancouver, April-May 2018), p. 1.

methods until the counterfeit money is completely indistinguishable from real money.[73] The goal of a GAN is to create image, voice or video generations that are undistinguishable from real audio-visual images. The speed, scale, and nuance of these two neural networks working together is much faster than what humans could achieve.[74] Therefore, GANs can be used to produce increasingly realistic deepfake images.

2.4 Benefits of deepfake technology

As mentioned earlier, deepfakes can be easily created and could serve all kinds of harmful purposes. They can be used to spread misinformation, misleading individuals into thinking negatively about those featured in the fake video.[75] Such videos can be used to erode trust in those featured in the video.[76] Deepfakes can also be used for financial scams.[77] There are countless other examples of how deepfake technology can be used to cause harm. Other such examples will be discussed extensively in section 2.5.

This section – however – will demonstrate that deepfake technology could also bring benefits to society or individuals. First of all, deepfake technology can have positive uses for art and self-expression. For example – using existing technology at the time – Peter Cushing was brought back to life in the 2016 Star Wars Movie "Rogue One."[78] As deepfake technology develops, it might be possible to change a dialogue in a movie without needing to reshoot an entire scene. Another example is a video of the Mona Lisa, where she moves her head, eyes, and mouth,[79] which shows that deepfakes can be used as art. Deepfakes can also be used by video artists to satirize or critique public figures or demonstrate a point.[80] They can

[73] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, 'Generative Adversarial Nets' (2014), p. 1 <https://arxiv.org/pdf/1406.2661.pdf> last accessed 11 November 2021.

[74] Ibid, p. 2-3.

[75] M. Masood, M. Nawaz, K.M. Malik, A. Javed, A. Irtaza, 'Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward' (2021), para. 2 <https://arxiv.org/ftp/arxiv/papers/2103/2103.00484.pdf>.

[76] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (2019) 107 California Law Review 1753, p. 1777.

[77] Ibid, p. 1772.

[78] D. Itzkoff, 'How 'Rogue One' Brought Back Familiar Faces' *The New York Times* (New York, 27 December 2016) <https://www.nytimes.com/2016/12/27/movies/how-rogue-one-brought-back-grand-moff-tarkin.html> last accessed 11 November 2021.

[79] T. Dafoe, 'Russian Researchers Used AI to Bring the Mona Lisa to Life and it Freaked Everyone Out' (2019) <https://news.artnet.com/art-world/mona-lisa-deepfake-video-1561600> last accessed 11 November 2021.

[80] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (2019) 107 California Law Review 1753, p. 1770.

be used to deliver a message. For example, a deepfake was created of David Beckham speaking in nine different languages, delivering an anti-malaria message in 2019.[81]

Secondly, deepfakes can be used for education or training purposes. Scientists at an online training platform called Udacity have been investigating whether deepfake technology can be used to automatically generate lecture videos from audio narration.[82]Additionally, deepfakes could very well be used in history classes, featuring deceased historic figures talking to students.[83] Start-ups like Synthesia have already developed technology for creating AI avatars that can be used in training videos.[84] These avatars can be created and customized and programmed to say anything in any language.[85]

Thirdly, deepfakes could have all sorts of individual benefits. For instance, individuals suffering from diseases like ALS could regain the ability to speak with their own voice.[86] Such technology could also be used for "digital storytelling,"[87] for example allowing individuals to interact with deceased relatives or relatives that live far away or are paralyzed. Besides that, deepfakes can be used to create digital replicas that can be used to consolidate individual legacies and keep them interactive.[88]

There are many other ways deepfake technology can be used for individual or societal benefit. They can for example be used when looking for a new pair of sunglasses online. One could create a deepfake of oneself and try on different models on a website.[89] Deepfakes can also be used to protect the identity of vulnerable individuals. In a documentary about anti-gay

[81] O. Oakes, ''Deepfake' voice tech used for good in David Beckham malaria campaign' (2019) <https://www.prweek.com/article/1581457/deepfake-voice-tech-used-good-david-beckham-malaria-campaign> last accessed 11 November 2021.

[82] 'Edtech company Udacity uses deepfake tech to create educational videos automatically' (2019) <https://www.fanaticalfuturist.com/2019/08/edtech-company-udacity-uses-deepfake-tech-to-create-educational-videos-automatically/> last accessed 11 November 2021.

[83] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (2019) 107 California Law Review 1753, p. 1769.

[84] E. Bond, 'AI Video Startup Synthesia Raises USD 12.5m for Multilingual Avatars' (2021) <https://slator.com/ma-and-funding/ai-video-startup-synthesia-raises-usd-12-5m-for-multilingual-avatars/> last accessed 11 November 2021.

[85] Ibid.

[86] S. Shakeri, 'Lyrebird Helps ALS Ice Bucket Challenge Co-Founder Pat Quinn Get His Voice Back' (2018) <https://www.huffingtonpost.ca/2018/04/14/lyrebird-helps-als-ice-bucket-challenge-co-founder-pat-quinn-get-his-voice-back_a_23411403/> last accessed 11 November 2021.

[87] N. Caporusso, 'Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology' in T. Ahram (ed.), *Advances in Artificial Intelligences, Software and Systems Engineering* (Springer 2020) p. 237.

[88] Ibid.

[89] The optician "Ace and Tate" already allows people to upload a picture of themselves to see how different glasses fit.

and anti-lesbian purges in Russia, the film producer used deepfake technology to mask the identity of those interviewed to protect them from harm.[90]

This section has made clear that one should not completely write off deepfake technology only because it can be used to cause harm to individuals. There are also benefits to the technology. However, to demonstrate the risks deepfake technology could pose, the next section will demonstrate that the technology can also be used to destroy people's lives.

2.5 Deepfakes used to intentionally harm to individuals

Although deepfake technology presents many new opportunities, many share the opinion that their risks outweigh their benefits.[91] In the past, only computer technology experts were able to convincingly manipulate audio-visual content, whilst now, anyone has access to AI technology that can be used to generate a deepfake. Simple deepfake applications can be installed on a mobile phone and are free of charge.[92] They can be used to generate a fake video within seconds. Given the ease of creating, posting, and sharing a deepfake online, deepfakes have great potential of spreading quickly and becoming increasingly pervasive in society.[93] Those featured in a falsified video may suffer many different harms as a result. This includes irreversible mental damage and even physical violence can be directed against them because the video they are featured in might cause public outcry.[94] Deepfake technology can be used to exploit individuals and sabotage their reputations.

Regarding exploitation, the technology could be used to extract some sort of (financial) benefit from others.[95] For instance, one could pretend to be someone's son in need of money and could create a deepfake of that son asking a parent for money and sending the clip to the parent (after getting a hold of the parent's phone number). Going a step further, one could even pretend to have kidnapped the son, demanding a ransom for his release. A deepfake of the kidnapped son could serve as proof that the son was really kidnapped. Instead of using a deepfake to mislead people, one could also extort others by threatening to release a deepfake

[90] J. Rothkopf, 'Deepfake Technology Enters the Documentary World' *The New York Times* (New York, 1 July 2020) <https://www.nytimes.com/2020/07/01/movies/deepfakes-documentary-welcome-to-chechnya.html> last accessed 11 November 2021.

[91] R. Chesney and D. Citron, 'Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?' (2018) <https://www.lawfareblog.com/deepfakes-looming-crisis-national-security-democracy-and-privacy> last accessed 11 November 2021.

[92] The app "WOMOBO" for instance, like mentioned earlier.

[93] J. Westling, 'Are Deep Fakes a Shallow Concern? A Critical Analysis of the Likely Societal Reaction to Deep Fakes' (2019), p. 5, available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3426174>.

[94] Ibid.

[95] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (2019) 107 California Law Review 1753, p. 1772.

of the victim, unless the victim does what he or she is told. A victim might be forced to provide bank account information or nude pictures of themselves (this last example is known as sextortion).[96] In this case, deepfakes are used as a leverage. Another way to use deepfake technology for exploitation purposes is by exploiting someone else's sexual identity for someone's own gratification.[97] Deepfake technology allows for faces, bodies, and voices to be swapped into real pornography.[98] The technology makes it possible for fans of a famous actress to create a pornographic deepfake of the actress for their own lust, at the expense of that actress.

Not all deepfakes are necessarily designed for financial benefit or sexual gratification. They can also be created simply to inflict pain. For example, once a relationship has ended, a frustrated ex-partner could create a violent deepfake sex-video, humiliating the other person. The ex-partner could also threaten to release such deepfake instead unless he or she receives intimate photos of the other person. Victims to such malpractices suffer violations of their "sexual privacy," as the barriers that protect information to their intimate lives are breached.[99] The consequences for victims can be devastating. They are reduced to sexual objects. One sextortion victim had the feeling she had been "virtually raped."[100] Others have been forced to change their names because the sexual-privacy invasions conducted against them were destructive to their identities.[101] Studies have shown that the emotional harms for such victims can be severe and lasting. Some victims have shown difficulty performing day-to-day activities. They experience anxiety and depression and some even contemplate suicide.[102]

In addition to exploitation or inflicting pain, deepfakes can be used to damage other people's reputation and sabotage their opportunities in life. This can be felt through any field of competition.[103] A deepfake of a successful bank employee stealing cash from a vault may cost that employee their job. A married person starred in a deepfake sex-video with another

[96] Ibid.

[97] See A. Dodge and E. Johnstone, 'Using Fake Video Technology to Perpetrate Intimate Partner Abuse' (2018), p. 6 at <https://withoutmyconsent.org/blog/2018-04-25-a-new-advisory-helps-domestic-violence-survivors-prevent-and-stop-deepfake-abuse/> for examples of individuals creating deepfake sex videos for their own gratification.

[98] D.K. Citron, 'Sexual Privacy' (2019) 128 Yale Law Journal 1870, p. 1921-1924.

[99] Ibid, p. 1874.

[100] P. Holley, 'The man who posed as his daughter's online boyfriend to get nude photos of her' *The Washington Post* (Washington, 17 March 2016) <https://www.washingtonpost.com/news/true-crime/wp/2016/03/17/the-man-who-posed-as-his-daughters-online-boyfriend-to-get-nude-photos-of-her/> last accessed 11 November 2021.

[101] D.K. Citron, 'Sexual Privacy' (2019) 128 Yale Law Journal 1870, p. 1925.

[102] Ibid, p. 1926. On page 1926-1927, Citron mentions examples of minors committing suicide after having become victims to sextortion.

[103] R. Chesney and D.K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (2019) 107 California Law Review 1753, p. 1774.

person can cost that person their marriage and could kill relationships with friends and family. A deepfake of a professional football athlete taking drugs or getting drunk could get that athlete kicked off the team. The examples of possibilities of using deepfake technology to damage other people's reputations are endless.

The reason for this is that deepfakes can sometimes be impossible to recognize as being fake.[104] Besides that, they have a potential of spreading to a large audience relatively quickly. This is because our world is connected through the internet, which facilitates the quick spread of "viral" content. Especially content that is "out of the ordinary" is likely to spread fast. Humans are more sensitive to negative information and information that stands out, which we want to share with other people.[105] One study showed that false stories spread ten times faster than real stories.[106] The study was conducted between 2006 and 2010. Data scientists had studied 126,000 Twitter news stories and classified them as true or false. The authors of the study assumed that the reason for false stories spreading faster was because false information was more "novel" and "evocative."[107] Next to this, humans easily pass along what other people think or post online, without verifying the information first. [108] This is because humans cannot know everything and we therefore rely on things other people say, assuming such information is correct.[109] Besides that, the spread of false information is exacerbated by "filter bubbles," that are the result of website algorithms predicting what information a user wants to see, irrespective of the information being true. [110] These algorithms are based on personalized searches and other information on the user. The result is that users no longer get to see information that contradicts what users think or believe.[111] Social media platforms are excellent information intermediaries, which allow for filter bubbles to form. This is because they allow users to re-share content, the algorithms highlight popular information read by

[104] For instance <https://www.youtube.com/watch?v=2rkQn-43ixs> shows how convincing a deepfake can be.

[105] R. Meyer, 'The Grim Conclusions of the Largest-Ever Study of Fake News' *The Atlantic* (Washington, 8 March 2018) <https://www.theatlantic.com/technology/archive/2018/03/largest-study-ever-fake-news-mit-twitter/555104/> last accessed 11 November 2021.

[106] S. Vosoughi, D. Roy, S. Aral, 'The spread of true and false news online' (2018) 359 Science 1146, p. 1148.

[107] Ibid, p. 1149.

[108] D. Easley, J. Kleinberg, *Networks, Crowds, and Markets: Reasoning about a Highly Connected World* (Cambridge University Press 2010) p. 503.

[109] Ibid.

[110] E. Bozdag, 'Bias in algorithmic filtering and personalization' (2013) 15 Ethics and Information Technology 209, p. 218.

[111] Ibid, p. 209.

"people like you"[112] and individuals can easily make connections with others that share their beliefs.[113]

It should therefore not come as a surprise that deepfakes are a powerful mechanism that can be used for the destruction of other people's reputation. Consider an employee of whom a deepfake sex-video has appeared online. The employer might worry that the employee's reputation might reflect badly on them and that the employee is distracted from their work.[114] Moreover, it might be impossible for the employee to prove that he or she has been targeted by a deepfake. This could result in the employee being fired. That employee will most likely also have trouble finding a new job afterwards. This is because most employers screen the online reputation of job applicants before hiring them.[115] One study by Reppler.com[116] – where 300 professionals involved in hiring processes were interviewed – showed that 91% of them used social networking sites to screen applicants. 69% had rejected an applicant based on information that was posted on social networking sites.[117] A job applicant that is featured in a harmful deepfake will not have the opportunity to explain that the video is fake. This is because they will often not even be contacted by recruiters, since employers rather hire people that have a better online reputation.[118]

This is just one example of how deepfake technology can be used for character assassination, sabotaging someone's career opportunities. I have not even touched upon the consequences for the employee's social life and all possible physiological problems he or she will be dealing with as a result. One should realize that deepfake technology can be used to bring harm to all areas of life.

2.6 Interim conclusion

This chapter has made clear that deepfake technology is the next big thing when it comes to content manipulation methods. It allows for the creation of fake audio-visual images of people that are indistinguishable from real audio-visual images. The technology is driven

[112] D.K. Citron, *Hate Crimes in Cyberspace* (2014 Harvard University Press), p. 67.

[113] N. Jankowicz, C. Oits, 'Facebook Groups are Destroying America' (2020) <https://www.wired.com/story/facebook-groups-are-destroying-america/> last accessed 11 November 2021.

[114] D.K. Citron, *Hate Crimes in Cyberspace* (2014 Harvard University Press), p. 7-8.

[115] Ibid.

[116] Reppler.com provided services for social media monitoring, which helped users to manage their online reputation.

[117] 'How Employers Use Social Media to Screen Applicants' <https://theundercoverrecruiter.com/infographic-how-recruiters-use-social-media-screen-applicants/> last accessed 11 November 2021.

[118] D.K. Citron, *Hate Crimes in Cyberspace* (2014 Harvard University Press), p. 8.

through deep learning algorithms, that are likely to become more efficient and produce more realistic content in the future.

Although there are several potential benefits of this technology in the fields of art, education, and autonomy, I believe the destructive potential of deepfakes outweigh these benefits. They can be used for exploitation purposes or to sabotage individuals in many different areas of competition. In essence, deepfakes can be used to destroy people's lives. The ease of them being created and spread and the difficulty to prove them being fake worries me. Thus, I hope this chapter has made clear that something must be done to regulate them.

I believe that deepfakes that are created after acquiring consent of the ones featured generally will not cause harm to individuals. At the same time, non-consensual deepfake sex-videos are likely to cause the most harm. Sometimes, a deepfake is created for art and self-expression purposes. In that case, the line between public benefit and individual harm becomes thin. The "drunk Nancy Poleski" clip shows us exactly that. On the one hand, the creator could have tried to prove a point by showing us how ridiculous Nancy Poleski's argument is, by making her seem drunk. On the other hand, the fake clip could disproportionately damage her reputation. The next chapter will dive deeper into the issues of non-consensual deepfakes, from a data protection perspective.

Chapter 3: Regulation through the GDPR

3.1 Introduction

This chapter will analyse how the GDPR applies to (harmful) deepfakes. It will become clear which main mechanisms can be relied upon, how they should be used to offer protection to individuals and whether these mechanisms are limited in effectiveness. In this way, this chapter is meant to provide for an understanding of to what extent the main mechanisms of the GDPR protect against the harms identified in the previous chapter. In so, is will become clear how the GDPR should be used to regulate harmful deepfakes, through its main mechanisms.

Section 3.2 will demonstrate to what extent personal data is being processed when a deepfake is created. Section 3.3 will highlight whether the GDPR applies to deepfakes in light of its material and territorial scope. Section 3.4 will explain the consequences of unlawful processing of personal data when a harmful deepfake is created. Section 3.5 will explain the situations where sensitive personal data is processed and the consequences thereof. Section 3.6 will describe different data subject rights and their limitations. Section 3.7 will explain the possibilities for victims of harmful deepfakes to seek compensation for damages they have suffered and who can be held liable for those damages. Section 3.8 will be used to answer the second sub-question.

3.2 Personal data

The first step in analysing to what extent the GDPR offers protection against the harmful use of deepfakes is identifying whether personal data is processed when creating a deepfake. If that is not the case, the GDPR does not apply.[119] The GDPR defines personal data as “any information relating to an identified or identifiable natural person.” The Article 29 Working Party (hereinafter “**WP29**”)[120] has guidelines available on how to define personal data[121] and breaks it up into four elements: (i) any information, (ii) relating to, (iii) identified or identifiable, (iv) natural persons.[122]

[119] Article 2(1) GDPR.

[120] The Article 29 Working Party was an independent European working party that dealt with privacy and data protection issues until the entry of the GDPR. It is the predecessor of the European Data Protection Board. See: ‘Article 29 Working Party,’ <https://edpb.europa.eu/about-edpb/more-about-edpb/article-29-working-party_en>, last accessed 11 November 2021.

[121] Article 29 Data Protection Working Party, ‘Opinion 4/2007 on the concept of personal data’ (2007) available at <https://www.clinicalstudydatarequest.com/Documents/Privacy-European-guidance.pdf> (hereinafter: “**WP 136**”).

[122] Ibid, p. 6.

Firstly, personal data must contain information. WP29 notes that the term "any information" shows the intent of the legislator to design the concept of personal data very broadly.[123] This line of reasoning can be derived from the *Nowak* Judgement,[124] in which the CJEU highlights that giving a wide scope to the term "any information" is appropriate, since ruling otherwise would result in certain information being excluded from the obligation to comply with data protection principles, safeguards and rights of data subjects.[125] Anything that is "information" can essentially fall under the concept of personal data, irrespective of its nature, content or format. This means that the nature of the information does not necessarily have to be true, the information does not have to concern private or family life and it does not matter in which format the information is kept.[126] This means that a deepfake contains information. Images as input data, also contain information. WP29 does not define what is exactly meant by "information." Apparently, it assumes the definition is self-evident.[127] Nadezhda Purtova notes that an argument can be made that everything can be or can contain information.[128]

Secondly, information must relate to an individual. According to WP29, that is generally the case if the information is about an individual. [129] It points out that information relates to an individual when a content, purpose or result element is present.[130] A content element is present when information is about a person. A purpose element is present when information is used with the purpose to evaluate an individual or treat an individual in a certain way or influence their status or behaviour.[131] A result element is present when information is likely to impact an individual such as them being treated differently as a result of processing of the data.[132] This result does not have to materialize. These elements should be considered as alternative conditions.[133] When creating a deepfake, input data consists of photos or audio-visual images of individuals. These data contain content elements, as the information is about individuals. An argument could be made that harmful deepfakes rather contain purpose elements since they are meant to influence the status of an individual in a negative way and

[123] Ibid.
[124] Case C-434/16 Peter Nowak v. Data Protection Commissioner [2017] ECLI:EU:C:2017:994.
[125] Ibid, para. 34.
[126] P. 7 WP 136.
[127] N. Purtova, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection law' (2018) 10(1) Law, Innovation and Technology 40, p. 48.
[128] Ibid, p. 50.
[129] P. 9 WP 136.
[130] Ibid.
[131] Ibid, p. 10.
[132] Ibid, p. 11.
[133] Ibid.

alternatively, result elements since they are likely to have a negative impact on a certain person's rights and interests. The outcome of this discussion is however irrelevant, since the three elements do not need to be assessed as cumulative conditions. WP29 notes specifically that "in particular, where the content element is present, there is no need for the other elements to be present to consider that the information relates to the individual."[134] In conclusion, at least a content element is present regarding the information used to create a deepfake. Whether the deepfake itself contains one of the elements depends on whether the deepfake resembles an actual person.

Thirdly, the information must relate to a natural person that is identified or identifiable. The GDPR only applies to natural persons that are living individuals. Information relating to deceased persons is not protected by the GDPR.[135] That is also the case for information which concerns legal persons.[136] WP29 notes that a natural person is identified when he or she can be distinguished within a group from other members of that group. A person is identifiable if he or she is not yet identified but it is possible to do so.[137] Consequently, the GDPR does not apply to anonymous data,[138] since it is not possible to identify an individual in that case. The GDPR notes that in order to determine whether an individual is identifiable, all the means that are "reasonably likely" to be used for the identification of that individual should be taken into account, such as singling out by either a controller or another person directly or indirectly.[139] Account should be taken of all objective factors that are likely to be used to identify an individual, such as the amount of time and costs required for identification and the available technology at the time of processing and the technological developments.[140] Moreover, in the *Breyer* case, the CJEU made clear that it is not required that all information necessary for the identification of a natural person are in the hands of one party. In that case, the CJEU essentially held that a dynamic IP address collected by an online media service provider constitutes personal data because the online media service provider has the means to legally obtain additional information from a third party internet service provider to identify a data subject.[141] In other words, the combining with the IP address and the necessary data for the identification of the data subject constitutes a means "likely reasonably to identify the data

[134] Ibid.
[135] Recital 27 GDPR.
[136] Recital 14 GDPR.
[137] P. 12 WP 136.
[138] Recital 26 GDPR.
[139] Ibid.
[140] Ibid.
[141] Case C-572/14 Patrick Breyer v. Bundesrepublik Deutschland [2016] ECLI:EU:C:2016:779, para. 47-48.

subject."[142] This means that information relating to a natural person can also be personal data, when a third party is necessary to identify that person. However, it must be noted that in the *Breyer* case, the CJEU implies also that the question whether something constitutes personal data must always be assessed on a case-by-case basis by stating that "if the identification of the data subject [by a third party] was prohibited by law or practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power (…) the risk of identification appears in reality to be insignificant."[143] This means that it is not sufficient that any third party worldwide could possibly identify an individual. When third party knowledge is required to identify an individual, it must constitute "a means likely reasonably to be used to identify the data subject"[144] and it must be legally and practically possible for the third party to identify the data subject.[145] Audio-visual images and pictures thus relate to identified or identifiable natural persons, if it is possible to identify the persons concerned, [146] based on the criteria discussed above.

Like described in section 2.3, the underlying mechanism for creating deepfakes consist of machine learning algorithms that swap face images and audio fragments of "target persons" into video image(s) of "source persons." Therefore, images are required of both the "target person" and the "source person", where their faces are visible. A face is in essence the fundamental identifying part of an individual. Whether an individual can be identified by a picture of their face is dependent on factors such as the quality of an image or the viewpoint.[147] Blurred faces or images of individuals in the distance are not likely to be considered personal data because an individual cannot be identified.[148] A clear picture of an individual's face that can be identified should however, be considered personal data.[149] Since algorithms are getting better at recognising faces online,[150] it is therefore relatively easy to trace a clear picture of one's face back to them. Applying face recognition technology

[142] Ibid, para. 45.
[143] Ibid, para. 46.
[144] Ibid, para. 45.
[145] Ibid, para. 46.
[146] See also: case C-212/13 František Ryneš v. Úřad pro ochranu osobních údajů [2014] ECLI:EU:C:2014:2428, para. 22.
[147] Article 29 Data Protection Working Party, 'Opinion 02/2012 on facial recognition in online and mobile services' (2012) available at <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2012/wp192_en.pdf> p.4.
[148] Ibid.
[149] Ibid.
[150] See for example: Y. Taigman, M. Yang, M.A. Razanto, L. Wolf, 'DeepFace: Closing the Gap to Human-Level Performance in Face Verification' (IEEE Conference on Computer Vision and Pattern Recognition, Colombus, September 2014), p. 1701-1708, where Facebook's face recognition technology is explained as being over 97% accurate of recognizing faces in a data set.

therefore can be regarded a "reasonably likely" way of identifying a natural person.[151] This means that a picture or video where someone's face is clearly visible can be regarded "information relating to an identified or identifiable natural person."[152] The input data for deepfakes should therefore be regarded personal data.

A question that now arises is whether the deepfake itself should be regarded personal data. Consider a deepfake of Donald Trump saying or doing things he has not. The deepfake itself should be regarded information – which is not necessarily true – that relates to Donald Trump – because it is about him – and Donald Trump is an easily identifiable natural person. In this case, the deepfake should therefore be regarded personal data. More interestingly, consider a deepfake amalgam of a "muscular Donald Trump," where images of Arnold Schwarzenegger and Donald Trump are used as input data. The deepfake should be regarded personal data, relating to Donald Trump but should Arnold Schwarzenegger also be regarded a data subject? If that is the case, it could raise difficult questions because both persons may have conflicting interests and views or opinions regarding the legitimacy of the video.[153] At first glance, it might seem more difficult to trace a picture of a body part back to an individual. In this case however, the deepfake contains body images of Arnold Schwarzenegger as an iconic individual, who people are likely to identify. Even when body images are used of individuals who are not famous, there is a chance that they will be recognized by people they know. Especially when a deepfake goes viral on a social media platform for instance,[154] it can be assumed that the chance that someone will recognize a body increases as more people are confronted with it. Chances that an individual's body is recognized could increase when birth marks, scars, or tattoos[155] are visible. Because a deepfake can spread quickly through social media, so can additional information on the identity of those featured. This way of identification should therefore be regarded "reasonably likely." The result is that pictures of body parts can relate to identified or identifiable natural person in many cases and thus,

[151] It must be noted however, that the Breyer case also states that a means is not "reasonably likely," when identification is prohibited by law (para. 46). At this point, facial recognition technology is not illegal but it is possible this could change in the future. See for instance: M. Heikkilä 'Europe's AI rules open door to mass use of facial recognition, critics warn' (2021) <https://www.politico.eu/article/eu-ai-artificial-intelligence-rules-facial-recognition/> last accessed 11 November 2021.

[152] Article 4(1) GDPR.

[153] B. van der Sloot, 'Editorial' (2020) 6(4) European Data Protection Law Review 477, pp. 478.

[154] See for instance: R. Meyer, 'The Grim Conclusions of the Largest-Ever Study of Fake News' *The Atlantic* (Washington, 8 March 2018) <https://www.theatlantic.com/technology/archive/2018/03/largest-study-ever-fake-news-mit-twitter/555104/> last accessed 11 November 2021.

[155] A. Jain and J. Lee, 'Scars, marks, and tattoos: a soft biometric for identifying suspects and victims' <https://spie.org/news/1282-scars-marks-and-tattoos-a-soft-biometric-for-identifying-suspects-and-victims?SSO=1> (2009) last accessed 11 November 2021.

personal data. In conclusion, Arnold Schwarzenegger should also be regarded a data subject regarding the "muscular Donald Trump" deepfake. In this example, both input data and output data are personal data. It must be noted that this does not mean that all deepfakes are in themselves personal data. There should always be a case-by-case assessment.[156] When creating a deepfake of a person that does not exist using different input data of thousands of different individuals, the deepfake does not relate to any identifiable natural person.[157] In this case, the deepfake itself is not personal data, although personal data is being processed when using input data of actual individuals to create it.

3.3 Scope and applicability

The GDPR protects natural persons "with regard to the processing of [their] personal data."[158] It applies to "the processing of personal data wholly or partly by automated means."[159] Processing activities include – among other things – collection, recording and alteration,[160] which are likely activities to be conducted when creating a deepfake. Because personal data is processed when a deepfake is created, it falls within the GDPR's material scope.

One important exception to the applicability of the GDPR is the "household exemption,"[161] which applies when individuals process personal data purely for "personal [use] or household activity." That is the case when the processing of personal data has "no connection to a professional or commercial activity."[162] A deepfake, which has been created for private use within a household circle, would at first glance seem to fall under this exception. However, when applying the *František Ryneš* judgement,[163] this conclusion is questionable: Mr. Ryneš had captured images through CCTV surveillance cameras of burglars breaking into his home. Although it was obvious that Mr. Ryneš used the image for his own private use, the Court of Justice of the EU (hereinafter: "**CJEU**") ruled that the household exemption did not apply because Mr. Ryneš had collected the images from a public domain. Apparently, the source of the personal data and not the use of it was determinative for the CJEU. If we apply this approach to a deepfake created for personal use, there is a risk that the

[156] Based on article 4(1) GDPR.
[157] As done on <thispersondoesnotexist.com>.
[158] Article 1(1) GDPR.
[159] Article 2(1) GDPR.
[160] Ibid.
[161] Article 2(2)(d) GDPR.
[162] Recital 18 GDPR.
[163] Case C-212/13 *František Ryneš v. Úřad pro ochranu osobních údajů* [2014] ECLI:EU:C:2014:2428.

household exemption does not apply, if personal data used to create the deepfake is collected from a public domain, like social media.[164] This means that the GDPR could be applicable to a deepfake that contains personal data even when a deepfake is created for personal use.

Another question is whether the household exemption applies when a deepfake is posted on a social media platform. This is because recital 18 of the GDPR states that "social networking and online activity undertaken within the context of [household] activities"[165] could fall within the household exemption. According to a guidance paper issued by WP29, information made "available to the world at large should be an important consideration when assessing whether or not processing is being done for personal purposes [but] this should not in itself be considered determinative."[166] This mean that the applicability of the household exemption should be assessed on a case-by-case basis. This case-by-case assessment was conducted in a case before a lower court in the Netherlands, where the court had to assess a claim of a mother of three under-aged children against the children's grandmother, who had posted pictures of her grandchildren online without consent.[167] The court noted that although it could not completely be ruled out that the household exemption applied, it had not been sufficiently established whether the grandmother had shielded off her account (to disallow an infinite number of people to have access to the pictures). [168] Besides that, it could not be established whether the picture could be available for third parties and search engines.[169] Accordingly – under these circumstances – the household exemption did not apply and the grandmother was ordered to remove the pictures. Although the court did not assess whether the grandmother had "shielded off" her account and in which situations an account can be considered to be "shielded off," apparently it can be an important consideration. However, when a harmful deepfake is published on social media, the intention of the creator will often be to spread it to a large audience.[170] Therefore, it is apparent that the household exemption will not apply in cases where a deepfake is published on a social media platform.

[164] Line of reasoning derived from B. van der Sloot, 'Editorial' (2020) 6(4) European Data Protection Law Review 477, pp. 478-479.

[165] Recital 18 GDPR.

[166] Article 29 Working Party, 'Annex 2 Proposals for Amendments regarding exemption for personal or household activities' (2013) available at < https://ec.europa.eu/justice/article-29/documentation/other-document/files/2013/20130227_statement_dp_annex2_en.pdf> p. 9. In case C-101/01 Criminal proceedings against Bodil Lindqvist [2003] ECLI:EU:C:2003:596, para 47, the CJEU made clear that the household exemption does not apply when personal data is published " on the internet so that those data are made accessible to an indefinite number of people.

[167] Rb. Gelderland 13 May 2020, ECLI:NL:RBGEL:2020:2521, *NJF* 2020/225.

[168] Ibid, para. 4.5.

[169] Ibid.

[170] Because this is the intention, I will not discuss what should be understood by "shielding off" one's account.

In addition to the material scope, the GDPR will only apply when the processing activities fall within the territorial scope. That is generally the case when controllers and processors have an establishment with activities in the EU,[171] irrespective of the location or nationality of the one whose data is being processed (hereinafter: "**data subject**"). [172] Although the term "establishment" is not defined in the GDPR, the recitals indicate that "establishment" implies the effective and real exercise of activity through stable arrangements."[173] The legal form of those arrangements is not decisive. It is apparent that a tendency exists to broadly apply these criteria to bring as much processing activities as possible within the scope of EU data protection legislation.[174] This was made clear in the *Google Spain* case, where the CJEU held that branches of subsidiaries based in the EU that do not carry out processing activities themselves can be inextricably linked to processing activities outside the EU, which will then fall under the territorial scope of the GDPR.[175] Because personal data is being processed when deepfakes are created, they fall under the territorial scope of the GDPR, when they can be linked to some sort of activity in the EU.[176]

Finally, creators have to be defined as "controller" or "processor." This is because the GDPR specifically applies to controllers and processors, where controllers are natural or legal persons that determine the purpose and means of data processing,[177] which is wide enough to include anyone who posts information about others online.[178] Processors are natural or legal persons that "process personal data on behalf of the controller."[179] Controllers and processors have to respect the different remedies and safeguards set out in the GDPR. An individual that creates a deepfake is the controller, since he or she determines why and how data is processed by for instance, collecting images from someone's social media account to create a deepfake

[171] Article 3(1) GDPR.
[172] European Data Protection Board, 'Guidelines 3/2018 on the territorial scope of the GDPR (Article 3) Version 2.1' (2019) available at < https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-32018-territorial-scope-gdpr-article-3-version_en>.
[173] Recital 22 GDPR. See also: Case C-230/14 Weltimmo s.r.o. v. Nemzeti Adatvédelmi és Információszabadság Hatóság [2015] ECLI:EU:C:2015:639, para. 31.
[174] 'Material and Territorial Scope: GDPR Series Part 1' (2016) <https://www.lexology.com/library/detail.aspx?g=5d778547-bc7e-42b2-acb2-2ec828d40a7d> last accessed 11 November 2021.
[175] Case C-131/12 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González [2014] ECLI:EU:C:2014:317, para. 60.
[176] This wide territorial scope of the GDPR has also led to the EU's protection of privacy rights spreading outside its boundaries, which is called the "Brussels effect." See for example: A. Bradford, 'The Brussels Effect' (2021) 107(1) Northwestern University Law Review 1, pp. 3 and 23. Bradford notes that since the DPD, over thirty countries have adopted EU-type privacy laws. In the meantime, countries are also adopting GDPR-like laws. California for example, has adopted the California Consumer Privacy Act.
[177] Article 4(7) GDPR.
[178] R. Wong, 'Social Networking: A Conceptual Analysis of a Data Controller' (2009) Vol. 14(5) Communications Law 142, p. 142.
[179] Article 4(8) GDPR.

(means) to destroy their reputation (purpose). Creators thus have to respect the rights and remedies in the GDPR.

At the same time, a search engine like Google can also be regarded a controller in relation to deepfakes. In the *Google Spain* case,[180] the CJEU found that the operator of a search engine determines the purposes and means of processing activities and it allows any internet user to make searches based on other people's names, "including to internet users who otherwise would not have found the web page on which those data are published."[181] If we follow this line of reasoning, a search engine should be regarded a controller, when it is possible to find a deepfake of an individual by using it. Additionally, social media platforms like Facebook should also be considered controllers, when a deepfake is published on their platform. This is because the platform's algorithm decides which information people will get to see (means) to make personalized recommendations for content and products users might be interested in (purpose).[182] Consequently, an algorithm could decide that people will get to see a deepfake and therefore disseminate[183] it, which is a processing activity. Social media platforms thus decide what the means and purposes of data processing are regarding deepfakes on their platforms. To what extent data subjects can exercise their rights vis-à-vis social media platforms and search engines will be explored in section 3.7.

3.4 Unlawful processing

In the previous sections it has been established that the GDPR applies to deepfakes. The result is that the creators of (harmful) deepfakes have to respect different data protection principles, rights and obligations. One of those principles is that personal data must be "processed lawfully, fairly and in a transparent manner."[184] Regarding a lawful ground, when creating a deepfake for a movie, processing activities may be necessary for performance of a contract,[185] provided the data is not sensitive data (section 3.5). In cases where someone creates a deepfake as satire or to criticize a public figure, processing activities might be "necessary for the purpose of the legitimate interests pursued by the controller or by a third

[180] Case C-131/12 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González [2014] ECLI:EU:C:2014:317.

[181] Ibid, para. 33. Displaying someone's personal data on a search results page can be regarded the means. On that same page, advertising links are displayed, which shows that personal data is processed in the context of commercial and advertising activity (purpose). See: para. 57.

[182] Facebook, 'Data Policy' <https://www.facebook.com/policy.php?ref=pf> last accessed 11 November 2021.

[183] Article 4(2) GDPR.

[184] Article 1(a) and 6(1) GDPR.

[185] B. van der Sloot, 'Editorial' (2020) 6(4) European Data Protection Law Review 477, p. 479 and art. 6(1b) GDPR.

party."[186] In those cases, little harm is done to the ones featured in the deepfake. However, when those interests are overridden by those of the data subject, this processing ground cannot be used.[187] The interests of data subjects are clearly overridden when data is processed to spread fake news, to commit identity fraud or when creating deepfake sex-videos.[188] This is because the creator has no legitimate interest in the creation of such video, whilst the victim potentially suffers a lot of harm.

Another lawful ground is consent. Consent must be "freely given, specific, informed and unambiguous."[189] In other words, a data subject must be "offered a genuine choice with regard to accepting or declining the terms offered (…),"[190] thus offered control over its decision. Consent can for instance be obtained from friends, who agree that a deepfake of them is created. However, consent cannot be relied upon when creating a harmful deepfake, since deepfakes created to harm the interests of others are usually created without the victim's knowledge and permission (why would someone consent to a harmful deepfake being created about them?). In this case, personal data will not be processed fairly or transparently either.

When data has been processed unlawfully to create a deepfake, data subjects have "the right to obtain from the controller the erasure of personal data concerning them.[191] If the deepfake itself is regarded personal data, the right to erasure will result in the deepfake getting removed. If only input data is personal data, only the input data will have to be removed. If the deepfake already has been created, the deepfake itself will not get removed after erasing input data.

However, when one has created a deepfake and has done so to exercise their "right of freedom of expression and information"[192] the GDPR does not require the creator to remove the deepfake or input data. The European Data Protection Board (hereinafter: "**EDPB**"), which replaced WP29 on 25 May 2018,[193] has issued guidelines on how to deal with the right to be forgotten of data subjects versus the right of freedom of expression and information of

[186] Article 6(f) GDPR.
[187] Ibid.
[188] B. van der Sloot, 'Editorial' (2020) 6(4) European Data Protection Law Review 477, p. 479.
[189] Article 4(11) GDPR.
[190] European Data Protection Board, 'Guidelines 05/2020 on consent under Regulation 2016/679 Version 1.1 (2020) available at <https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf>.
[191] Article 17(1)(d) GDPR.
[192] Article 17(3)(a) GDPR.
[193] Article 94(2) GDPR.

other parties.[194] It has noted that a balancing exercise between these rights must be conducted. Essentially, this balancing exercise is about weighing data subject privacy rights against rights of interested parties, like internet users, and the freedom of expression, which includes free access to information.[195] According to the CJEU in *Google Spain*, this balancing exercise depends, "in specific cases, on the nature of the information in question and its sensitivity for the data subject's private life and on the interest of the public in having that information, an interest which may vary, in particular, according to the role played by the data subject in public life."[196]

If thousands of images are collected from social media without permission to generate images of individuals that do not exist, like done on the website "thispersondoesnotexist.com," "freedom of expression" could be used as an exception. In that case, a balancing exercise must be conducted between the data subject's privacy rights and the interest of the public having the information. In this specific case, the fact that personal data has been collected from social media to create the generated pictures does not seem to influence the life of the data subjects in a negative way. Therefore, the fact that no lawful processing ground applies does not require the controller to erase the data collected on the data subject according to the GDPR.

A similar conclusion could follow from such a balancing exercise when – for instance – a satirical deepfake is published by a politician to mock a political rival. In that case, the deepfake is created to bring across a political message. If the deepfake is obviously fake, the interest of the public having access to it could outweigh the damage done to the individual featured.

It should be obvious that in many of the situations described in chapter 2, this balancing exercise will shift into the direction of favouring protection to the rights of data subjects. A deepfake sex-video distributed online – for instance – will potentially cause a lot of damage to a data subject's private life and there is no interest in the public in having the information. In this specific case, the victim can exercise his or her right of erasure of the personal data, resulting in the deepfake itself to get removed. The creator of the deepfake will not be able to rely on their freedom of expression and freedom of information.

[194] European Data Protection Board, 'Guidelines 5/2019 on the criteria of the Right to be Forgotten in the search engines cases under the GDPR (part 1) Version 2.0' (2020) available at <https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-52019-criteria-right-be-forgotten-search-engines_en>.

[195] Ibid, p. 11.

[196] Case C-131/12 *Google Spain SL, Google Inc. v. AEPD, Mario Costeja González* [2014] ECLI:EU:C:2014:317, para. 81. See also for instance: Case C-136/17 *GC et al. v. CNIL* [2019] ECLI:EU:C:2019:773, para. 66.

3.5 Special categories of personal data

Now that it is clear which legal grounds can be used to create a (harmful) deepfake, this section will explore to what extent deepfakes could process special categories of personal data, also called "sensitive data," which is prohibited in principle under the GDPR. Special categories of personal data under the GDPR are personal data "revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation."[197]

Since harmful deepfakes could feature anyone saying or doing anything, one could imagine that in many cases, they will contain such sensitive data. Returning to the example of Donald Trump saying or doing things he has not, let us assume the things he says strongly contradict the ideals of the United States Republican party. Because the deepfake contains information – that not necessarily is true – relating to Donald Trump, the deepfake itself should be regarded personal data. Furthermore, the deepfake contains sensitive personal data revealing political opinions. This type of processing is in principle prohibited under the GDPR.[198] This could also be the case when a deepfake shows someone engaged in criminal activities[199] or explicit sexual content.[200]

The GDPR does however, give several exceptions to the processing prohibition of sensitive personal data.[201] It seems improbable that any of these grounds can be invoked when creating a harmful deepfake, with the exception when the "processing relates to personal data which are manifestly made public by the data subject."[202] When the deepfake of Donald Trump has been created using data collected from his social media accounts, this exception seems to apply. However – because the deepfake itself can be regarded personal data – this line of reasoning is not very convincing. Although personal data is processed when images are collected[203] from Donald Trump's social media account, the creation of the deepfake and making it publicly available should be considered new processing activities. In this case, the political deepfake video of Donald Trump is entirely new, meaning that it was not made

[197] Article 9(1) GDPR.
[198] Ibid.
[199] Recital 75 GDPR.
[200] B. van der Sloot, 'Editorial' (2020) 6(4) European Data Protection Law Review 477, p. 450.
[201] Article 9(2) GDPR.
[202] Article 9(2)(e) GDPR.
[203] Article 4(2) GDPR.

publicly available in the past. Consequently, this exception cannot be successfully invoked in this case.

In most cases, explicit content shown in a deepfake is what will make the deepfake harmful. This explicit content can be regarded sensitive data if it is for example of sexual nature and will often not have been made publicly available by a data subject for the creation of a deepfake. Therefore, the creation of non-consensual deepfakes with such explicit content is in principle prohibited under the GDPR.

3.6 Other data subject rights and their limitations

Apart from data subjects' right to have their data erased, various other data subject rights under the GDPR are applicable when it comes to the creation of deepfakes.[204] The right to be informed requires controllers to ensure that data subjects are properly informed in "easily accessibly form, using clear and plain language"[205] that their data is processed. Data subjects thereby have the right to access the information that is being processed concerning them.[206] Data subjects have the right to have inaccurate personal data rectified.[207] Data subjects may exercise the right to restrict personal data processing, when they contest the accuracy of the personal data[208] or the processing activities are unlawful. Finally, data subjects could exercise their right to object[209] to data processing if it is done so "for the purpose of the legitimate interests pursued by the controller or by a third party."[210] A creator of a harmful deepfake must respect these rights. Moreover, social media platforms and search engines – as controllers – will also be obliged to comply with these data subject rights. For instance, a victim of a harmful deepfake could request Facebook to remove a deepfake, because the creator of the deepfake has unlawfully processed the data to create it.[211]

When a harmful deepfake appears online, an individual featured in it thus has a vast number of possibilities to do something about it. Although the GDPR presents data subjects with many different remedies, there are some serious limitations that influence their effectivity in a negative way. First of all, the creator of a harmful deepfake video is not likely

[204] I do not discuss the right to data portability and the right not to be subject to automated decision making because I do not deem these rights relevant when it comes to harmful deepfakes.

[205] Article 12(1) GDPR.

[206] Article 15(1) GDPR.

[207] Article 16 GDPR.

[208] Article 18(1)(a) GDPR.

[209] Article 21(1) GDPR.

[210] Article 6(1)(f) GDPR.

[211] Article 17(1)(d) GDPR.

to inform the victim[212] that their data is being processed to create the deepfake. The result is that the victim might not know the video exists or will discover about it after damage has already been done. Moreover, the victim might not be able to exercise their data subject rights directly against the creator, since it might be impossible to know where the deepfake came from and who published it online. This is because some platforms allow for anonymous use[213] and deepfake creators could make their IP addresses untraceable, using special software.[214] Besides that, why would the person that has created the deepfake respect data protection rights in the first place?

Victims could perhaps alternatively exercise their rights before a social media platform or search engine. The problem here is that the victim might be fighting a running battle, as a new deepfake could easily reappear[215] (somewhere else online) and it might be a time-consuming process before the it is removed. However – as the next section will demonstrate – social media platforms and search engines are the most suitable actors before whom data subjects could exercise their data subject rights.

3.7 Liability, the right to compensation and joint controllership

The GDPR offers several mechanisms that allow victims of harmful deepfakes to counter violations by controllers or processors. Most importantly, victims can hold controllers liable for damages suffered because of GDPR infringements.[216] Like with exercising data subject rights, it will be difficult for victims of harmful deepfakes to exercise their right to compensation because they will need to prove that there is an unlawful act and more importantly, establish who can be defined as the controller.[217]

Can a victim alternatively seek compensation from a search engine or social media platform? In section 3.3 it was already pointed out that search engines and social media platforms should be regarded controllers and thus have obligations to respect data subject

[212] D. Harris, ‘Deepfakes: False pornography is here and the law cannot protect you’ (2019) 17 Duke Law & Technology Review 99, p. 112.

[213] R. Delfino, ‘Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn’s Next Tragic Act’ (2019), 88 Fordham Law Review 887, p. 899.

[214] A. Greenberg, ‘It’s About To Get Even Easier to Hide on the Dark Web’ (2017) <https://www.wired.com/2017/01/get-even-easier-hide-dark-web/> last accessed 11 November 2021.

[215] See for instance: B. Clark, ‘Pornhub promissed to ban ‘deepfakes’ Videos. And it failed miserably’ <https://thenextweb.com/news/pornhub-promised-to-ban-deepfakes-videos-and-it-failed-miserably> (last accessed 11 November 2021) which shows that some platforms have difficulty stopping people from (re)posting harmful deepfakes.

[216] Article 82 GDPR.

[217] Van Alsenoy, B ‘Liability under EU Data Protection Law: From Directive 95/46 to the General Data Protection Directive’ (2016), 7 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 271, p. 273-274.

rights under the GDPR. The *Google Spain* case illustrated that search engines can be held liable for damages of data subjects if they fail to remove inadequate, irrelevant, or excessive information.[218] Therefore, a victim can seek compensation from a search engine if it does not respect their right to be forgotten.[219] This line of reasoning can be extended to social media platforms. If a deepfake – which is inaccurate information – is published on such platform, data subjects can exercise the right to be forgotten if information is published unlawfully,[220] which will be the case if a harmful deepfake is published. If the social media platform then does not respect the right to be forgotten, it could be held liable for damages by data subjects.[221]

This analysis shows that data subjects could in principle seek compensation from a search engine or social media platform if they fail to respect their right to be forgotten by not removing harmful deepfakes. The problem here is that it could take time before that right is respected, meaning the damage could already have been done. However – since many deepfakes spread through social media platforms[222] – this problem could be dealt with if social media platforms can be held liable under the GDPR for the mere publication[223] of a deepfake itself. If that is the case, it could mean that the GDPR (indirectly) imposes content moderation obligations on social media platforms.[224]

The GDPR gives data subject the power to exercise their rights against each joint controller,[225] including the right to compensation and liability.[226] The question is thus specifically, whether social media platforms can be regarded joint controllers with creators of harmful deepfakes. That is the case if these two actors determine jointly the purposes and

[218] Case C-131/12 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González [2014] ECLI:EU:C:2014:317, para. 70 & 92.

[219] Article 82(1) GDPR.

[220] Article 17(1)(d) GDPR.

[221] Article 82(1) GDPR.

[222] Given the ease of spreading misinformation through social media platforms. See M.A Britt, J.F. Rouet, D. Blaum & K. Millis, 'A Reasoned Approach to Dealing With Fake News' Vol. 6(1) Policy Insights from the Behavioral and Brain Sciences 94, p. 99.

[223] G. de Gregorio, 'The e-Commerce Directive and GDPR: Towards Convergence of Legal Regimes in the Algorithmic Society?' (2019) Robert Schuman Centre for Advanced Studies Research Paper No. RSCAS 2019/36 available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3393557> p. 9.

[224] In the literature it has been noted that the GDPR encourages the removal of content that is inadequate, irrelevant or excessive, which provides a way to moderate content. See C. Castets-Renard, 'Algorithmic content moderation on social media in EU law: illusion of perfect enforcement' (2020) 2020(2) Journal of Law, Technology & Policy 283, p. 314. This implies content moderation in "reactive sense," where I investigate whether the GDPR imposes content moderation obligations in "preventive sense."

[225] Article 26(3) GDPR.

[226] Article 82(1) GDPR.

means of processing activities.[227] The assessment of joint controllership should be a factual analysis of the actual influence parties have on the means and purposes of processing.[228]

The relevant case law on this topic illustrates that the scope of joint controllership is broad.[229] First of all, the CJEU made an interesting remark about the possibility of joint controllership stemming from "technical configurations" in the *Google Spain* case.[230] In the underlying case, the CJEU determined that Google is a controller because it – among other things – indexes personal data and makes it available to internet users.[231] Even though the CJEU did not directly deal with issues revolved around joint controllership, it noted that "even if [the] option for publishers of websites [to opt out from Google's indexing] were to mean that they determine the means of that processing jointly with [Google], this finding would not remove any of the latter's responsibility."[232] Although these remarks were made for a purely hypothetical situation, apparently it is possible for joint controllership to arise between Google and websites when those websites use (or do not use) certain technical settings.[233]

Additionally, the CJEU acknowledged the possibility of loose relationships in the *Wirtschaftsakademie* judgement,[234] where the CJEU held that the administrator of a Facebook page is a joint controller with Facebook. In its assessment on whether the administrator and Facebook jointly determined the means and purposes of processing, the CJEU focused on how the settings of the page would technically allow Facebook to collect personal data from users in the group.[235] It noted that an administrator of a Facebook page "gives Facebook the opportunity to place cookies on the computer or other device of a person visiting its fan page,

[227] Article 26(1) GDPR. See also European Data Protection Board, 'Guidelines 07/2020 on the concepts of controller and processor in the GDPR Version 1.0' (2020) available athttps://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202007_controllerprocessor_en.pdf> p. 17.

[228] Ibid.

[229] J. Chen, L. Edwards, L. Urquhart, D. McAuley, 'Who Is Responsible for Data Processing in Smart Homes? Reconsidering Joint Controllership and the Household Exemption' (2019) Vol. 10(4) International Data Privacy Law 279, p. 284.

[230] Case C-131/12 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González [2014] ECLI:EU:C:2014:317.

[231] Ibid, para. 28.

[232] Ibid, para. 40.

[233] J. Chen, L. Edwards, L. Urquhart, D. McAuley, 'Who Is Responsible for Data Processing in Smart Homes? Reconsidering Joint Controllership and the Household Exemption' (2019) Vol. 10(4) International Data Privacy Law 279, p. 283.

[234] Case C-210/16 Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v. Wirtschaftsakademie Schleswig-Holstein GmbH [2018] ECLI:EU:C:2018:388.

[235] J. Chen, L. Edwards, L. Urquhart, D. McAuley, 'Who Is Responsible for Data Processing in Smart Homes? Reconsidering Joint Controllership and the Household Exemption' (2019) Vol. 10(4) International Data Privacy Law 279, p. 283.

whether or not that person has a Facebook account."[236] An administrator "contributes to the processing of personal data of visitors to its page [by Facebook]" because it "has influence on the processing of personal data [by Facebook] through the possibility of defining "the criteria in accordance with which the statistics are to be drawn up and even designat[ing] the categories of persons whose personal data is to be made use of by Facebook."[237] In other words, Facebook page administrators can be held to jointly determine the "means" of processing with Facebook because they technically allow Facebook to process personal data of page visitors and also the "purpose" by taking part[238] in determining the purposes by "adding their own aims into the mix."[239] It is important to highlight that in this case, the CJEU departs from the doctrine that only actors that determine the reasons and ends for data processing activities are controllers.[240] Instead of looking at Facebook's general purposes and means, the CJEU assesses the individual processing activities within the system.[241] It was however left to the national court to determine the level of responsibility that resulted from joint-controllership.[242]

Moreover, the scope of joint controllership was broadened even further in the is the *Jehovan todistajat* case,[243] where the CJEU had to give an opinion on whether a the Jehova's Witness Community and its members should be regarded joint controllers because personal data is collected through door-to-door preaching. The CJEU made clear that it is not necessary to use written guidelines to determine the purposes of data processing[244] and the involved party does not need to have access to the personal data in question.[245]

[236] Case C-210/16 Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v. Wirtschaftsakademie Schleswig-Holstein GmbH [2018] ECLI:EU:C:2018:388, para. 35.

[237] Ibid, para. 36.

[238] Ibid, para.. 39.

[239] J. Knibbe, 'Complying with the GDPR on social media – interactions' (2020) <https://www.linkedin.com/pulse/complying-gdpr-social-media-interactions-jorren-knibbe> last accessed 11 November 2021. These purposes are the managing and promoting of activities according to para. 39 of the *Wirtschaftsakademie* judgement.

[240] See for instance the SWIFT case: Article 29 Working Party, Opinion 10/2006 on the processing of personal data by the Society for Worldwide Interbank Financial Telecommunication (SWIFT)' (2006) available at <https://www.dataprotection.ro/servlet/ViewDocument?id=234>.

[241] R. Mahieu, J. van Hoboken & H. Asghari, 'Responsibility for Data Protection in a Networked World' (2019) Vol. 10(1) Journal of Intellectual Property, Information Technology and Electronic Commerce Law 84, p. 89.

[242] R. Mahieu & J. van Hoboken, 'Fashion-ID: Introducing a phase-oriented approach to data protection?' (2019) <https://europeanlawblog.eu/2019/09/30/fashion-id-introducing-a-phase-oriented-approach-to-data-protection/> last accessed 11 November 2021. See also: J. Chen, L. Edwards, L. Urquhart, D. McAuley, 'Who Is Responsible for Data Processing in Smart Homes? Reconsidering Joint Controllership and the Household Exemption' (2019) Vol. 10(4) International Data Privacy Law 279, p. 291 where it is noted that the existence of joint responsibility does not imply equal responsibility regarding different processing activities.

[243] Case C-25/17 Jehovan todistajat [2018] ECLI:EU:C:2018:551.

[244] Ibid, para. 67.

[245] Ibid, para. 69.

Finally, the CJEU confirmed that joint controllership can arise without any legal relationship between parties concerned in the *Fashion ID* case,[246] in which the CJEU had to clarify whether websites that use the Facebook "like button"– giving Facebook access to certain information on the website – would result in such websites becoming joint controllers with Facebook. The CJEU held that placing the "like button" on its website, Fashion ID exerts decisive influence over how personal data of visitors is processed, which would not have occurred without it,[247] resulting in Facebook and Fashion ID jointly determining the "means" of processing.[248] The "purpose" of processing are also jointly determined, since the "processing operations are performed in the economic interests of both Fashion ID and Facebook."[249] However, the CJEU also distinguishes the different stages of processing for which Fashion ID can be regarded controller, influencing for which processing activities it can be held liable for GDPR infringements (the phase-oriented approach).[250] Taking the overall "chain" of processing activities into account, the CJEU notes that Fashion ID can be regarded controller for the *collection* and *disclosure* of personal data through the Facebook 'Like' button[251] because it jointly determines the means (by implementing the Facebook like button and thus exerting a "decisive influence" over data processing of personal data of visitors)[252] and purposes (optimalisation of publicity of Fashion ID's goods and services)[253] of that personal data. That is not the case however, for the subsequent processing operations of the collected personal data by Facebook.[254] The result of this, is that joint controllers do not always share equal responsibilities regarding specific processing activities. Mara Paun notes that the CJEU also seems to introduce a knowledge element into the equation, requiring Fashion ID to be "fully aware" of data processing, for it to be able to be held responsible.[255] She notes however, that it is unclear whether this qualification will be accepted as a condition in future case law.[256]

[246] Case C-40/17 Fashion ID GmbH & Co. [2019] ECLI:EU:C:2019:629.
[247] Ibid, para. 78.
[248] Ibid, para. 79
[249] Ibid, para. 80.
[250] Ibid, para. 70-85. The phase-oriented approach divides data processing into different, separate phases.
[251] Ibid, para. 84.
[252] Ibid, para. 75-78.
[253] Ibid, para. 80-83.
[254] Ibid, para. 76. Fashion ID can thus be held liable only for the (i) collection and (ii) transmission of personal data and not the subsequent processing activities by Facebook.
[255] M. Paun, 'On the Way to Effective and Complete Protection (?): Some Remarks on Fashion ID' (2020), Vol. 9(1) Journal of European Consumer and Market Law 35, p. 37.
[256] Ibid.

The abovementioned case law stresses the wide scope of joint controllership, although the existence of it does not imply equal responsibility for different processing activities.[257] For joint controllership to arise between creators of harmful deepfakes and social media platforms, it must be established whether the means and purposes of specific processing activities are jointly determined by these actors.[258] When creators of harmful deepfakes publish them online (means), they will often do so with the goal of deliberately harming other individuals by spreading it to a large audience (purposes). At the same time, a social media platform's algorithm will disseminate it (means) to other platform users with the purpose of managing and promoting activities (purpose).[259] Such platform will thus have an interest in content spreading to a large audience. For these reasons, an argument can be made for social media platforms "contribut[ing] to the processing of personal data," [260] since it influences the audience that will be able to see the deepfake through its algorithm (means). Besides that, the social media platform could be said to have contributed to the purposes of processing activities because it also has an interest in online content spreading to a larger audience. If this line of argumentation is true, the creator of a harmful deepfake and a social media platform should be regarded joint controllers.[261]

A question that remains is whether such a social media platform can be held responsible by victims,[262] in light of the phase-oriented approach introduced in the *Fashion ID* case. In the underlying case, the CJEU treats Fashion ID as a joint controller regarding only the collection and disclosure of personal data to Facebook, because these are the operations where it *actually* determines the means and purposes. According to Paul de Hert and Georgios Bouchagiar, this line of reasoning implies that Facebook – as the other joint controller – can escape liability, because it does not *actually* determine the means and purposes of the

[257] Primoz Gorkic notes that the result is that the scope of data protection has immensely expanded as a result. See P. Gorkic, 'Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW e.V.: More Control, More Data Protection for Website Visitors?' (2019) 5 European Data Protection Law Review 579, p. 581. On the other hand, AG Bobek warned in his conclusion in Fashion ID that having a number of persons co-responsible could result in in fact no one being responsible. See Case C-40/17 Fashion ID GmbH & Co. KG v Verbraucherzentrale NRWeV (Opinion of Advocate General, 19 December 2018), paras. 75 and 92 and M. Paun, 'On the Way to Effective and Complete Protection (?): Some Remarks on Fashion ID' (2020), Vol. 9(1) Journal of European Consumer and Market Law 35, p. 37.

[258] Case C-40/17 Fashion ID GmbH & Co. [2019] ECLI:EU:C:2019:629, para. 74.

[259] Case C-210/16 Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v. Wirtschaftsakademie Schleswig-Holstein GmbH [2018] ECLI:EU:C:2018:388, para. 39.

[260] Ibid, paras 36 and 39.

[261] It must be noted that it is not evident how the CJEU would consider this line of reasoning. It could, for instance, also distinguish a social media platform's economic interest vis-à-vis the creator's intention to produce harm, resulting in no joint controllership to arise. On the other hand, it could consider these actors joint controllers under the principle of "effective and complete protection" as used in para. 60 of the Google Spain case.

[262] Article 82(1) GDPR.

collection and disclosure of personal data.[263] In my opinion, this is a logical reasoning, because the CJEU considered in *Fashion ID* that (joint) controllers only have responsibility for processing activities where they *actually* determine the means and purposes. If we apply this line of reasoning to a harmful deepfake being *published* on a social media platform by a creator and the subsequent *dissemination* of it by the platform's algorithm, we need to consider these two separate processing activities. Although the platform itself could be regarded the means of processing (it provides a space for publication), an argument could be made that it does not determine the purpose. Platforms enjoy economic benefit[264] when content is disseminated but not for the mere publication, per se and the dissemination of content should be considered a separate processing activity.[265] If this is true, the result is that social media platforms cannot be held responsible for the publication, even if it can be regarded a joint controller.[266] In that case, I believe that the GDPR does not oblige them to moderate illegal content in a preventive way.[267] This line of reasoning could extend to search engines.[268]

[263] P. de Hert & G. Bouchagiar, 'Fashion ID and decisively influencing Facebook plugins: a fair approach to single and joint controllership' (2021) Vol. 7(27) Brussels Privacy Hub Working Paper, available at <https://euagenda.eu/publications/fashion-id-and-decisively-influencing-facebook-plugins-a-fair-approach-to-single-and-joint-controllership>, p. 13. If in some way, Facebook would *actually* determine the means and purposes of collection and disclosure of the personal data (collected and disclosed by Fashion ID), it could be held liable.

[264] Case C-40/17 Fashion ID GmbH & Co. [2019] ECLI:EU:C:2019:629, para. 80.

[265] To my opinion, the first processing activity for which social media platforms *actually* determine the means and purposes of processing is dissemination of the deepfake through the platform's algorithm.

[266] Also, because it might not have knowledge of its unlawful publication. See: M. Paun, 'On the Way to Effective and Complete Protection (?): Some Remarks on Fashion ID' (2020), Vol. 9(1) Journal of European Consumer and Market Law 35, p. 37.

[267] If this line of reasoning is not accepted by the CJEU, a social media platform could try to escape liability through art. 82(3) GDPR. Besides that, in para. 70 of the *Fashion ID* case, the CJEU holds that "all relevant circumstances" matter when assessing the degree of liability. When taking all circumstances into consideration, I do not believe a social media platform can be held liable for every single GDPR-infringement of individuals who publish content on their platform.

[268] Although I did not make an assessment regarding search engines, the outcome will be the same. This is due to the "phase-oriented approach" introduced in the Fashion ID case, like discussed.

Nevertheless, social media platforms engage in content moderation anyways.[269] Moreover, social media platforms and search engines should react when they have been made aware that certain information impacts a data subject adversely, according to *Google Spain*.[270]

3.8 Interim conclusion

In this chapter, it has become clear how the GDPR's main mechanisms should be used to protect individuals against the harmful use of deepfake technology. The GDPR is applicable to deepfakes since input data is personal data. In many cases deepfakes are in themselves, personal data as well. Furthermore, sensitive data is often processed when harmful deepfakes are created because, for instance, data concerns a natural person's sexual orientation. The processing of sensitive data is in principle prohibited.

The consequence of the applicability of the GDPR to deepfakes is that creators – as controllers – must respect various data subject rights, like the right to be informed, the rights to rectification, to erasure, to compensation and to be forgotten. These are important mechanisms that protect against the harmful use of deepfake technology. However, the effectiveness of these mechanisms is seriously limited since creators are highly unlikely to respect data subject rights and data processing principles because of their intention to cause harm. Moreover, it might be unclear for a victim to know who the creator is and where the deepfake came from and therefore, they do not know before whom they should exercise their rights.

On the other hand, victims of harmful deepfakes are not left completely empty-handed as they can exercise the right to be forgotten before social media platforms and search engines and can hold them liable if that right is not respected, but it could take a while before content is removed and by the time it is, the damage will have already been done.

Therefore, an important question that has been addressed in this chapter is whether social media platforms can be held liable under the GDPR for the mere fact that a deepfake is

[269] That does not matter since different social media platforms engage in content moderation. See 'Social Media: Misinformation and Content Moderation Issues for Congress' (2021) <https://crsreports.congress.gov/product/pdf/R/R46662> p. 6. For example, Twitter and Facebook provide lists of inappropriate content at <https://business.twitter.com/en/help/ads-policies/ads-content-policies.html> and <https://transparency.fb.com/en-gb/policies/>. As a result, harmful deepfakes (like deepfake porn videos) are prohibited and will be removed on these platforms. Moreover, in 2016 a code of conduct was agreed between the EU and data companies like Facebook, Google, Twitter, YouTube and Microsoft, where the importance of adhering to EU legislation was highlighted and these companies were obliged to respond within 24 hours after a notification of a breach. See: *Parliamentary Papers II*, 2018/19, 35 080, 3, pp. 12-13.

[270] B. van Alsenoy, 'Liability under EU Data Protection Law: From Directive 95/46 to the General Data Protection Directive' (2016), 7 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 271, p. 277.

published on their platform. If that is the case, the GDPR (indirectly) imposes preventive content moderation obligations on them. An argument has been made that social media platforms (and in the same manner search engines) could be regarded joint controllers with creators of harmful deepfakes. However, the *Fashion ID* case (and various case notes in the literature) suggests that the adopted "phase-oriented approach" limits responsibility of joint controllers to processing activities where they *actually* determine the means and purposes. Social media platforms and search engines will not *actually* determine the means and purposes of the creation of a harmful deepfake, and an argument has been made that this is also the case for its publication,[271] which would mean that the GDPR does not impose preventive content moderation obligations on them.

The conclusion of this chapter is that the GDPR offers various mechanisms that protect against the harmful use of deepfake technology. Exercising the "right to be forgotten" before social media platforms and search engines seems to be the most effective and important mechanism. In turn, these are the most suitable actors to turn to for protection. However, the effectiveness of the right to be forgotten (and other remedies) is limited because of its ex post nature, meaning it will be exercised once the damage has already been done. Therefore, the GDPR does not fully protect natural persons against the harmful use of deepfake technology. The next chapter will explore how other regulatory frameworks such as criminal law and targeted legislation could be used to address this regulatory issue.

[271] If this is not accepted by the CJEU, social media platforms could escape liability through art. 82(3) GDPR or para. 70 of *Fashion ID*. I believe this also results in the GDPR not imposing content moderation obligations on them.

Chapter 4: GDPR in relation to other regulatory mechanisms

4.1 Introduction

In the previous chapter, I demonstrated that the right to be forgotten is the main mechanism of the GDPR that should be relied upon when protecting natural persons against harmful deepfakes. I have also shown that the GDPR does not provide effective ex ante remedies. It does not obligate social media platforms and search engines to prevent publications of harmful deepfakes in the first place. This shortcoming limits the GDPR's effectiveness. meaning that many problems described in chapter 2 are not effectively dealt with by the GDPR. Therefore, it is important to consider whether other regulatory frameworks like criminal law and targeted legislation could be used to address these shortcomings. This chapter will demonstrate how the GDPR should relate to these other regulatory mechanisms. In so, this chapter will demonstrate how the GDPR should be used to regulate the harmful use of deepfake technology, in light of other regulatory frameworks like criminal law and targeted legislation.

Section 4.2 will explore whether the GDPR should be looked at as a solution to regulate harmful deepfakes in the first place and how it should relate to other regulatory solutions. Section 4.3 will then be used to assess how criminal law could address shortcomings of the GDPR when regulating the harmful use of deepfake technology. Section 4.4 will demonstrate how targeted legislation could be used to address remaining shortcomings. Section 4.5 will be used to answer the third sub-question.

4.2 Role of the GDPR considering other regulatory solutions

This section will explore whether the GDPR should be used to regulate harmful deepfakes in the first place and if so, what its role should be compared to other regulatory frameworks. This approach is therefore different than materially applying the provisions of the GDPR to deepfakes as done in chapter 2. This section will rather be an overarching analysis of whether the GDPR is an appropriate mechanism to regulate harmful deepfakes and what its role should be considering other regulatory solutions.

The text of the GDPR was adopted in 2016,[272] whilst the first deepfake appeared in 2017.[273] It can therefore be said that the drafters of the GDPR were not aware that the instrument would be used to regulate deepfakes in the future. Nevertheless, recital 15 of the GDPR indicates that the GDPR was designed to be technology neutral. This indicates that even though the GDPR was not explicitly designed to regulate deepfakes directly, it was nevertheless meant to regulate such technology. Moreover, using the GDPR to regulate harmful deepfakes ensures protection to the rights and interests of natural persons and it therefore responds to the CJEU's call to ensure "effective and complete protection of data subjects."[274]

However, some authors question whether applying the GDPR to certain technological developments is the right approach. For example, Rania El-Gazzar and Karen Stendal note that the characteristics of certain AI technologies are challenged by the GDPR.[275] Because of the autonomy of AI systems, they lead to compliance issues regarding the accountability principle.[276] AI systems may produce discriminatory results and thus provide biased representations of reality, which is not in line with the fairness principle.[277] When machine learning techniques are used, data processing purposes might be ambiguous, clashing with the purpose limitation principle.[278] According to these authors, these characteristics require a different approach for regulating AI-systems, than the GDPR.[279]

Although the GDPR poses certain challenges for AI systems, it has many benefits as well, as demonstrated in chapter 3. It has therefore also led to many positive reactions in the literature. Jan Philipp Albrecht, for instance, notes that the GDPR's harmonizing character will provide certainty and coherence and create a level playing field for all companies within

[272] Article 99 GDPR.

[273] See for example: B. Goggin, 'From porn to 'Game of Thrones': How deepfakes and realistic-looking fake videos hit big' (2019) <https://www.businessinsider.com/deepfakes-explained-the-rise-of-fake-realistic-videos-online-2019-6?IR=T> last accessed 11 November 2021.

[274] Case C-131/12 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González [2014] ECLI:EU:C:2014:317, para. 60.

[275] R. El-Gazzar & K. Stendal, 'Examining How GDPR Challenges Emerging Technologies' (2020) Vol. 10 Journal of Information Policy 237, pp. 266-267.

[276] Ibid, p. 263.

[277] Ibid, p. 264.

[278] Ibid, pp. 264-265. There are also clashes with other principles such as the data minimization principle.

[279] Ibid, p. 262. Tal Zarsky holds a similar position regarding big data technologies and notes that the GDPR undermines the ability to exercise big data analytics. See T. Zarsky, 'Incompatible: The GDPR in the Age of Big Data' (2017) Vol. 47, No. 4(2) Seton Hall Law Review 995, p. 1004. Additionally, a white paper by the Blockchain Bundesverband notes that the GDPR is outdated, since it does not account for Blockchain's decentralized technologies. It provides recommendations and interpretations of the law in that respect. See: N. Eichler, S. Jongerius, G. McMullen, O. Naegele, L. Steininger & K. Wagner, ´Blockchain, data protection, and the GDPR' <https://www.bundesblock.de/wp-content/uploads/2019/01/GDPR_Position_Paper_v1.0.pdf> last accessed 11 November 2021.

the EU, benefitting business and consumers.[280] Moreover, Jesper Zerlang, considers the GDPR's open wording as one of the biggest benefits of the instrument. He notes that it was developed with the future in mind, motivating companies to secure their systems and avoid data breaches[281] and should be regarded a major step forward for the privacy of EU citizens.[282]

If personal data has the potential to impact individuals, some form of legal protection should be triggered.[283] The GDPR grants individuals such protection. Although in can be argued that its approach poses certain challenges to new technologies that run on AI, I believe it should nevertheless be used to regulate the harmful use of deepfake technology. This is because it was designed to be technology neutral[284] and thus was indirectly designed to regulate technology like deepfakes. Moreover, it offers safeguards to natural persons, protecting their rights and interests.[285] Although there are limitations to the GDPR's effectiveness in regulating harmful deepfakes because of its ex post character, I believe these limitations should not be dealt with by applying other regulatory frameworks alternatively but rather complementary. In this way, the rights and interests of natural persons are best ensured.[286] In the next sections, I will analyze how criminal law and targeted legislation could complement the GDPR by addressing its main shortcoming that it does not provide effective ex ante protection.

4.3 Criminal law as a regulatory solution

This section will set out to what extent criminal law could be used as a complementary regulatory framework to deal with issues revolved around the harmful use of deepfake

[280] J.P. Albrecht, 'How the GDPR Will Change the World' (2016) 2 European Data Protection Legislation Review 287, p. 288.

[281] J. Zerlang, 'GDPR: a milestone in convergence for cyber-security and compliance' (2017) Vol. 6 Network Security 8, p. 8.

[282] T. Baxevani, 'GDPR Overview' (2019), p. 4, available at <https://www.researchgate.net/publication/333560686_GDPR_Overview>. On the other hand, this step forward for privacy has also resulted in a step backward for other important topics, like public health. Elizabeth Gourd notes that limitations in data sharing have disrupted international health research projects by prohibiting the transfer of data outside the EU. See: E. Gourd, 'GDPR obstructs cancer research data sharing' (2021) Vol. 22(5) The Lancet Oncology 592, p. 592.

[283] N. Purtova 'The law of everything. Broad concept of personal data and future of EU data protection law' (2018) Vol. 10 Law, Innovation and Technology 40, p. 74.

[284] Recital 1 GDPR.

[285] Most importantly, the right to be forgotten as laid down in article 17 GDPR.

[286] Looking at criminal law, for example, as an alternative framework would mean that the GDPR's right to be forgotten would no longer be used. This is a undesirable consequence, because enacting the right to be forgotten before social media platforms and search engines is a strong remedy for individuals.

technology. The rationale of criminal law is, among other things, (i) justice for victims,[287] (ii) reconciliation for offenders[288] and (iii) retribution and deterrence for society.[289]

The doctrine of state sovereignty limits the scope and applicability of criminal law to crimes that have been committed on the territory of a state.[290] The scope of applicability of criminal law is thus narrower than the GDPR. Moreover – because criminal law is not harmonized within the EU[291] – it plays out differently in different (European) countries regarding harmful deepfakes. This can be problematic because crimes that are committed online ignore national borders.[292] Because Dutch criminal law has many provisions that can be used as examples to show how the harmful use of deepfake technology could be regulated, I will use it as a starting point. Moreover – since I am Dutch – I also have a good understanding of Dutch criminal law.

Which provisions in the Dutch Criminal Code (hereinafter: "**DCC**") are applicable depends on the circumstances of the case. When someone spreads a harmful deepfake of someone else with the purpose of harming their reputation or good name, this will be qualified as aggravated[293] defamation.[294] This will, for instance, be the case when a deepfake appears of a professional football athlete taking drugs and drinking, which has been spread with the purpose to get that player kicked off the team. When there is no intention to harm someone's reputation or good name, a harmful deepfake could also qualify as insult.[295] This could for instance be the case, if someone were to create a deepfake of a stutterer, stuttering much more heavily than they really would and sending it in a WhatsApp group with that person in it.[296] A deepfake could also be created for (financial) gain. When one creates a deepfake of someone else asking for money for instance and proceeds to send it to that

[287] E. Maculan & A. Gil, 'The Rationale and Purposes of Criminal Law and Punishment in Transitional Contexts' (2020) Vol. 40(1) 132, p. 142.

[288] Ibid, p. 152.

[289] Ibid, p. 142. For a more in-depth outline of the aims and functions of criminal law, see A. Ashworth & J. Horder *Principles of Criminal Law* (seventh edition, Oxford University Press 2013), p. 16.

[290] M. El Zeidy, 'The Principle of Complementarity: A New Machinery to Implement International Criminal Law' (2002) 23 Michigin Journal of International Law 869, p. 870.

[291] Barring harmonization of provisions in the cybercrime convention. See: Council of Europe, 'Convention on Cybercrime' [2003] available at <https://www.europarl.europa.eu/meetdocs/2014_2019/documents/libe/dv/7_conv_budapest_/7_conv_budapest_en.pdf> (hereinafter: "**CCC**").

[292] E.J. Koops, 'The Internet and its Opportunities for Cybercrime' Vol. 1 Transnational Criminology Manual (2011) 735, p. 735.

[293] In the Netherlands defamation (article 261 Dutch Criminal Code (hereinafter: "**DCC**")) is aggravated when a perpetrator knows that their claim about a victim is contrary to the truth.

[294] Article 262 DCC. See also: S. van der Hof, Wraakporno op internet – een verkenning van de (on)mogelijkehden voor een strafrechtelijke aanpak' (2016) 65 Ars Aequi 54, p. 56.

[295] Article 266 DCC.

[296] Example derived from HR 11 February 2014, ECLI:NL:HR:2014:306, where a suspect imitated a stutterer in an elevator, in the presence of other people. This was qualified as "insult" by the Supreme Court.

person's parents, this will be classified as fraud.[297] Such a deepfake could also be used for coercion[298] or menace.[299]

Because these provisions are quite general, they also regulate pornographic deepfakes. However, there are other provisions in the DCC which could be applicable to deepfakes in specific cases like the provisions on "revenge porn"[300] and "child pornography."[301] Revenge porn and child pornography are criminalized in the Netherlands because of the pressing social need.[302] When it comes to child pornography, article 240b DCC applies to realistic sexual depictions of non-existent persons that are seemingly under eighteen years.[303] This is because modern computer technology is able to produce realistic images that cannot or hardly can be distinguished from real.[304] If these realistic depictions were not criminalized, this would lead to problems in courts where it would always have to be proven that a real minor was involved.[305] The result is that the mere possession[306] of a deepfake of child pornography falls under this definition. Regarding revenge porn, the legislator did not specifically criminalize "realistic sexual depictions" that are not real. Article 139h DCC only mentions "images." However, in the explanatory report, the legislator mentions that this provision has been designed to encompass future versions of revenge porn.[307] Therefore, an argument can be made that a pornographic deepfake used for revenge porn would fall under its scope because it is created without consent and published to harm an individual. However, in this situation, it is necessary to establish an intention by the perpetrator to bring harm to the victim when publishing the deepfake, which could be difficult to prove, especially when such a deepfake is shared online without any additional comments.[308]

In certain cases, a public prosecutor can order providers of electronic communications services (like WhatsApp) to block access to certain information to end or prevent a criminal

[297] Article 326 DCC.

[298] Article 284 DCC.

[299] Article 285 DCC.

[300] Article 139h DCC.

[301] Article 240b DCC. See also article 9 CCC and Lanzarote Convention, CETS 201.

[302] Regarding revenge porn, the Dutch legislator found it necessary to specifically criminalize it because it believed revenge porn can deeply affect the lives of people in a negative way. See: *Parliamentary Papers II,* 2018/19, 35 080, 3, p.4. Regarding child pornography, the Dutch legislator found child pornography to be a very serious act, and therefore wanted it to be specifically criminalized. See: *Parliamentary Papers II* 1994/95, 23 682, 5*,* p. 13.

[303] *Parliamentary Papers II,* 2000/01, 27 745, 6, p. 8.

[304] *Parliamentary Papers II*, 2001/01, 27 745, 3, p. 10.

[305] Ibid, p. 6.

[306] Article 240b DCC also criminalizes distribution, offering, publicly displaying, producing, importing, conveying in transit, exporting or accessing child pornography.

[307] *Parliamentary Papers II*, 2018/19, 35 080, 3, p. 5.

[308] M.L.R. Goudsmit, 'Criminalising Image-based Sexual Abuse: an Analysis of the Dutch Bill against Revenge Pornography' (2019) 68 Ars Aequi, pp. 445-446.

act.[309] The effect of this regarding harmful deepfakes is limited, because of the ease that the deepfake will reappear[310] and the fact that the damage will already have been done because the deepfake has already been widely spread.

The DCC provides various solutions regarding the harmful use of deepfake technology. The provisions are technology neutral, thus applicable to this new technology. To put it even more strongly, the provisions discussed are actually technology independent, since they abstract completely away from technology.[311] Moreover, criminal law provides an ex ante direction to members of society (besides ex post assessments of violations) because of deterrence, whereas the right to be forgotten as the most important remedy of the GDPR is rather of ex post nature.[312] These advantages show that criminal law is a suitable instrument to regulate harmful deepfake technology. There are however, also disadvantages. The scope and applicability of national criminal laws is narrower than that of the GDPR, which could be said to limit its effectiveness.[313] Besides that, criminal law does not provide solutions for victims in finding out where a deepfake came from and quickly responding before damage is done. It also requires it to be made clear who created the deepfake – which can be difficult[314] – whilst the GDPR gives data subjects the possibility to exercise the right to be forgotten before social media platforms and search engines, not requiring a creator to be identified.

In the end, using criminal law to regulate deepfakes has advantages and disadvantages compared to using the GDPR. However, it is not necessary to "choose" one of either regulatory framework. These frameworks should complement each other. In this way, victims are best protected. Victims can rely on both criminal law's deterrent character and GDPR's right to be forgotten with respect to social media platforms and search engines. An issue that

[309] Article 125p Dutch Code of Criminal Procedure (hereinafter: **"CCP"**). Applies only to coercion, menace, fraud and child pornography according to article 67(1) CCP.

[310] See for instance: B. Clark, 'Pornhub promissed to ban 'deepfakes' Videos. And it failed miserably' <https://thenextweb.com/news/pornhub-promised-to-ban-deepfakes-videos-and-it-failed-miserably> (last accessed 11 November 2021) which shows that some platforms have difficulty stopping people from (re)posting harmful deepfakes.

[311] E.J Koops, M. Lips, C. Prins & M. Schellekens, 'Should ICT Regulation Be Technology neutral?' (2006) Vol. 9 IT & Law Series, p. 5 available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=918746>.

[312] Although the GDPR also has some deterrent mechanisms in the shape of administrative fines (article 83 GDPR). These can for instance be imposed on social media platforms and search engines when they fail to respect the right to be forgotten. In this way, these actors can be encouraged to respect this right, in fear of the consequences of fines.

[313] This is because the doctrine of state sovereignty limits the scope and applicability of criminal law to crimes that have been committed on the territory of a state, whilst harmful deepfakes are mostly published in cyberspace, which has no borders.

[314] Certain platforms allow for anonymous use: R. Delfino, 'Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act' (2019), 88 Fordham Law Review 887, p. 899. Moreover, creators could make their IP addresses untraceable: A. Greenberg, 'It's About To Get Even Easier to Hide on the Dark Web' (2017) <https://www.wired.com/2017/01/get-even-easier-hide-dark-web/> last accessed 11 November 2021.

remains however, is that they still do not prevent deepfakes from appearing and causing damage in the first place.

4.4 Need for specific regulation?

In the previous sections it has been illustrated that the GDPR and criminal law are suitable legislative instruments that could complement each other to regulate the harmful use of deepfake technology. However, there are serious flaws that limit their effectiveness. Most importantly, these instruments do not prevent deepfakes from appearing or damage from being done. Moreover, it takes time before the exercise of remedies have effect and those remedies do not prevent content from reappearing. It is improbable that other existing areas of law offer solutions for these problems and thus some authors suggest that current laws are inadequate to effectively deal with harmful deepfakes.[315] I agree with these authors. Existing laws do not prevent harmful deepfakes from appearing.

Because the remedies in the GDPR and criminal law are limited in effectiveness, the question that now will be turned to is whether specific legislation for regulating harmful deepfake technology is more appropriate than using the GDPR and criminal as existing regulatory frameworks. Regulating the technology itself could serve effective to protect the rights of individuals.[316] However, we must ask ourselves whether it is the technology itself that threatens these rights or whether those rights are under threat because of non-technological harmful behaviour.[317] Like explained in section 2.4, deepfake technology can be used for purposes that are not necessarily harmful. Therefore, it does not seem desirable that the technology itself is regulated through – for instance – an all-out ban. Although there are "technological risks"[318] that deepfakes are used for harmful purposes, the technology itself

[315] See for instance D. Harris, 'Deepfakes: False pornography is here and the law cannot protect you' (2019) 17 Duke Law & Technology Review 99, where the author explains how there are not sufficient legal remedies in place to effectively prevent harmful pornographic deepfakes and A. P. Gieseke, 'The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography' (2020) 73 Vanderbilt Law Review 1479, p. 1500 or E. Meskys, J. Kalpokiene, P. Jurcys, A. Liaudanskas, 'Regulating Deep-Fakes: Legal and Ethical Considerations' (2019) Vol. 15(1) Journal of Intellectual Property Law & Practice 24, pp. 27-28 where the authors highlight limitations in copyright law and tort law. Other authors argue that existing laws are sufficient to regulate harmful deepfakes. See for instance D. Greene, 'We Don't Need Lew Laws for Faked Videos, We Already Have Them' (2018) <https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them> last accessed 11 November 2021.

[316] T. Prosser, *The Regulatory Enterprise: Government Regulation and Legitimacy* (Oxford University Press 2010), pp. 13-15.

[317] L.B. Moses, 'Regulating in the Face of Sociotechnical Change' in R. Brownsword, E. Scotford & K. Yeung (eds) *The Oxford Handbook of Law, Regulation and Technology* (Oxford University Press 2017), p. 6, available at <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199680832.001.0001/oxfordhb-9780199680832-e-49>.

[318] This term is used by Eileen Fischer in E. Fischer, *Risk regulation and Administrative Constitutionalism* (Hart Publishing 2007).

does not increase the risk of harms being done to individuals per se. It rather might lead to activities that could lead to harms.[319] The technology itself is therefore not harmful but rather certain activities where it is used for.

Lyria Benett Moses notes that the solution to problems revolved around new technologies in many cases can be dealt with in a technology neutral way. In taking this approach, the regulatory regime can deal with underlying problems instead of focussing on the means through which they arise.[320] For instance, if an offensive act like defamation can be accomplished by using different technologies, it makes more sense to criminalize the act under a technology neutral provision instead of adopting different offences for specific technologies.[321] When it comes to harmful deepfakes, the "regulatory rationale"[322] is preventing that they are used for harmful purposes. In that respect, deepfakes are merely an example of a new technology that generates the same problems. Following this line of reasoning, not the technology itself but rather the effects should be regulated.[323] The GDPR and criminal law already regulate these effects. New, specific legislation regulating these effects therefore is redundant.

These instruments are however, limited in effectiveness. The question is whether law should be looked at in the first place to deal with these gaps. Ronald Leenes notes that we should not regulate "just because we can" but could also "let the market handle things."[324] Even though the GDPR does not impose content moderation obligations on social media platforms and search engines – which would result in more effective protection for individuals – these actors often engage in such activities in accordance with their codes of conduct.[325] In

[319] Argument derived from L.B. Moses, p. 8.
[320] Ibid, p. 13. Moses notes that in other cases specific legislation may be the best means of achieving certain regulatory goals. Whether specific or neutral regulation should be applied should be assessed on a case-by-case basis.
[321] S.W. Brenner, *Law in an Era of 'Smart' Technology* (Oxford University Press 2007), p. 13.
[322] Moses, p. 13.
[323] R. Leenes, E. Palmerini, E.J. Koops, A. Bertolini, P. Salvini & F. Lucivero, 'Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues' (2017), Vol. 9(1) 1, p. 7.
[324] R. Leenes, 'Regulating New Technologies in Times of Change' in L. Reins, *Regulating New Technologies in Uncertain Times* (T.M.C. Asser Press 2019), p. 7.
[325] Different social media platforms engage in content moderation. See 'Social Media: Misinformation and Content Moderation Issues for Congress' (2021) <https://crsreports.congress.gov/product/pdf/R/R46662> p. 6. For example, Twitter and Facebook provide lists of inappropriate content at <https://business.twitter.com/en/help/ads-policies/ads-content-policies.html> and <https://transparency.fb.com/en-gb/policies/>. As a result, harmful deepfakes (like deepfake porn videos) are prohibited and will be removed on these platforms. Moreover, in 2016 a code of conduct was agreed between the EU and data companies like Facebook, Google, Twitter, YouTube and Microsoft, where the importance of adhering to EU legislation was highlighted and these companies were obliged to respond within 24 hours after a notification of a breach. See: *Parliamentary Papers II*, 2018/19, 3, pp. 12-13. It must be noted however, that pats of these codes of conduct might be influenced by specific EU regulations that impose content moderation

this way, problems revolved around the GDPR being ineffective in dealing with the problems at the source can be nuanced. New laws are not necessarily required.[326]

If we do choose to look at legislative solutions for the harmful use of deepfake technology, the question is how should we regulate? In the draft proposal for regulating AI, creators of deepfakes are obligated to disclose that content is artificially generated or manipulated, when they create a deepfake.[327] However, this approach does not seem to bring us one step closer to preventing harmful deepfakes from appearing in the first place. Even with these transparency obligations, I do not believe creators will be stopped in publishing harmful content.

Social media platforms and search engines could rather be targeted in new legislation. They could be required to (further) develop technologies to detect and filter harmful deepfakes.[328] One downside to this however, is the potential "chilling effect" of the freedom of expression by individuals.[329] If for instance, social media platforms block profiles of individuals who disseminate a harmful deepfake online and subsequently, report them to authorities, individuals could stop sharing videos of political figures in general out of fear of sharing a deepfake unknowingly.[330] Therefore, a balance of interest is required between the dangers harmful deepfakes present and the concerns of potential chilling effects on freedom of speech.[331] That being said, I believe the dangers of harmful deepfakes far outweigh these concerns[332] because of their destructive potential.[333]

Another downside is that it might be burdensome and costly for social media platforms and search engines to develop technology on top of what they already do to detect malicious

obligations on platforms regarding for instance terrorist content and child sexual abuse material. For an overview of the different regulations see: A. De Streel et al. 'Online Platforms' Moderation of Illegal Content Online' (2020) available at <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf>.

[326] On the other hand, it can be argued that the threat of new regulation(s) encourages preemptive self-regulation by big tech platforms, which helps protect the interests of individuals. See for instance: M. Cusumano, A. Gawer, D. Yoffie, 'Can Self-Regulation Save Digital Platforms?' (2021) available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3900137>.

[327] Article 52(3) Proposal for Artificial Intelligence Act.

[328] E. Meskys, J. Kalpokiene, P. Jurcys, A. Liaudanskas, 'Regulating Deep-Fakes: Legal and Ethical Considerations' (2019) Vol. 15(1) Journal of Intellectual Property Law & Practice 24, p. 30.

[329] J.W. Penney, 'Internet surveillance, regulation, and chilling effects online: a comparative case study' (2017) Vol. 6(2) Internet Policy Review 1, p. 2.

[330] Line of reasoning derived from A. Pasetski, 'Deepfakes: A New Content Category for a Digital Age' (2020) Vol. 29(2) William & Marry Bill of Rights Journal 503, p. 528 & 531.

[331] Ibid, p. 4.

[332] See for example: R. Chesney and D. Citron, 'Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?' (2018) <https://www.lawfareblog.com/deepfakes-looming-crisis-national-security-democracy-and-privacy> last accessed 11 November 2021.

[333] However, I would also like to express my doubts regarding this kind of technology, since it seems that it could lead to situations where deepfakes are also "filtered" which are not necessarily harmful.

content.[334] Besides that, the technology behind deepfakes is becoming more advanced, which might make it more difficult for AI to detect them,[335] resulting in a cat and mouse game between social media platforms and search engines on the one hand and creators on the other. As the technology develops or new technologies appear, new issues will emerge.[336] Although, imposing obligations on social media platforms and search engines probably is the most effective way to deal with issues the GDPR and criminal law cannot deal with in the short term, it is thus ineffective in the long term.

4.5 Interim conclusion

This chapter was dedicated to exploring how regulatory solutions such as criminal law and targeted legislation could be used to address shortcomings of the GDPR when it comes to regulating harmful deepfakes. I began my assessment by arguing that the GDPR was meant to regulate harmful deepfakes, because of its technology neutral design and its goal to protect the rights and interests of individuals, which are under threat by harmful deepfakes. Therefore, the GDPR should be used to regulate harmful deepfakes and should be complemented by other regulatory frameworks to better protect the rights and interests of individuals. I then explained how criminal law and targeted legislation could be used as such complementary solutions to address shortcomings of the GDPR

The main advantage of using criminal law is that it in some sense has ex ante effect because of its deterrent character. However, one of the disadvantages is the lack of remedies for victims to quickly respond when a harmful deepfake does appear online. The perpetrator would need to be identified before they could be prosecuted. Instead, the GDPR offers much faster solutions to victims through the right to be forgotten. Criminal law could however be used to identify and prosecute creators of harmful deepfakes in a later stage. In this way, criminal law could complement the GDPR. In the end, criminal law will not effectively prevent harmful deepfakes from appearing online.

Unfortunately, other existing areas of law share this same limitation. I therefore assessed whether specific legislation could be used to deal with this regulatory gap. Specific legislation to regulate the harmful use of deepfake technology, targeted at social media platforms and

[334] Inspired by S. Brenner, 'Distributed Security: A new Model of Law Enforcement' (2005), p. 40, available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=845085>.

[335] See for instance: L. Kelion, 'Deepfake detection tool unveiled by Microsoft' *BBC News* (London, 1 September 2020) <https://www.bbc.com/news/technology-53984114> last accessed 11 November 2021.

[336] M. Goodwin, 'A Dimensions Approach to Technology Regulation' in M.E.A. Goodwin, E.J. Koops & R.E. Leenes, *Dimensions of technology regulation* (Wolf Legal Publishers 2010), p. 1.

search engines is the most effective approach in dealing with remaining issues the GDPR and criminal law cannot solve. In this way, targeted legislation could be used to address shortcomings of the GDPR when regulating harmful deepfakes in the short term. However, as deepfake technology develops, targeted legislation will be ineffective in the long term. Moreover, it might be burdensome and costly for social media platforms and search engines to develop content moderation technology on top of what they already do to detect malicious content and there is a risk such targeted legislation has a chilling effect on freedom of speech.

The conclusion of this chapter is that the GDPR should be complemented by other regulatory frameworks such as criminal law and targeted legislation, to protect individuals against the harmful use of deepfake technology. Considering these other mechanisms, the GDPR should be used to quickly respond to the publication of harmful deepfakes on social media platforms and search engines, whilst criminal law could be used for deterrence and to prosecute individuals at a later stage. Although there are some serious disadvantages to targeted legislation, it could be used to prevent harmful deepfakes from appearing in the first place.

Chapter 5: Conclusion

5.1 Gap in the literature

Rapid developments in the field of AI now make it possible to create convincing audio-visual images of individuals saying or doing things they have not. Although this technology brings positive benefits to society, many authors agree that their risks outweigh these benefits.[337] Deepfakes can be used to sabotage or exploit individuals[338] and pose great risk to society as a whole.[339] Given the ease of creating, posting and sharing a deepfake online, they have great potential of spreading quickly and becoming increasingly pervasive.[340]

Different solutions have been proposed in the literature. Examples include regulation through data protection law,[341] privacy law,[342] criminal law,[343] civil law,[344] competition law[345] and intellectual property law.[346] Soft law mechanisms like training,[347] corporate policies and voluntary actions by social media firms[348] have also been proposed.

Nonetheless, substantial research on the role of the GDPR for the regulation of harmful deepfakes is limited. Some sources mention the GDPR, but they lack extensive review on how its remedies should be used and whether they are limited in effectiveness.[349] Moreover,

[337] R. Chesney and D. Citron, ‘Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?’ (2018) <https://www.lawfareblog.com/deepfakes-looming-crisis-national-security-democracy-and-privacy> last accessed 11 November 2021.

[338] R. Chesney and D.K. Citron, ‘Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security’ (2019) 107 California Law Review 1753, p. 1772.

[339] M.B. Kugler and C.L. Pace, ‘Deepfake Privacy: Attitudes and Regulation’ (2021) 116 Northwestern University Law Review, p. 11, available at: <https://ssrn.com/abstract=3781968>.

[340] J. Westling, ‘Are Deep Fakes a Shallow Concern? A Critical Analysis of the Likely Societal Reaction to Deep Fakes’ (2019), p. 5, available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3426174>.

[341] A. Eigbedion, ‘Deepfakes: Legal & Regulatory Considerations in Nigeria’ (2020), p 7-9, available at: <https://ssrn.com/abstract=3670644>.

[342] K. Farish, ‘Do Deepfakes Pose a Golden Opportunity? Considering Whether English Law Should Adopt California’s Publicity Right in the Age of Deepfake (2020) Vol. 15, No. 1 Journal of Intellectual Property Law & Practice 40, p. 44-46.

[343] R. Delfino, ‘Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn’s Next Tragic Act’ (2019), 88 Fordham Law Review 887, p. 926-928.

[344] R. Chesney and D.K. Citron, ‘Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (2019) 107 California Law Review 1753, p. 1792.

[345] A. Eigbedion, ‘Deepfakes: Legal & Regulatory Considerations in Nigeria’ (2020), p 11-12, available at: <https://ssrn.com/abstract=3670644>.

[346] E. Meskys, J. Kalpokiene, P. Jurcys, A. Liaudanskas, ‘Regulating Deep-Fakes: Legal and Ethical Considerations’ (2019) Vol. 15(1) Journal of Intellectual Property Law & Practice 24, p. 27.

[347] M. Westerlund, ‘The Emergence of Deepfake Technology: A Review’ (2019) Vol. 9(11) Technology Innovation Management Review 40.

[348] O. Schwartz, ‘Deepfakes aren’t a tech problem. They’re a power problem’ (2019) <https://www.theguardian.com/commentisfree/2019/jun/24/deepfakes-facebook-silicon-valley-responsibility> last accessed 11 November 2021.

[349] Of the cited papers, only the one written by K. Farish mentions GDPR. However, this is limited to a brief mentioning of two challenges for asserting one’s GDPR rights regarding deepfakes. The GDPR is assessed in

there are no academic sources out there that explore how the GDPR should be used to regulate harmful deepfakes, considering other regulatory frameworks.

The gap in the literature which I aim to fill with my research thus is an extensive analysis of the role the GDPR should play in regulating the harmful use of deepfake technology. Conducting this research is important to an overall discussion on how deepfakes can and should be regulated.

5.2 Main research question

The main question that my thesis revolved around answering is:

"How should the GDPR be used to offer protection against the harmful use of deepfake technology?"

5.3 Findings

My research has confirmed the GDPR's wide territorial reach and applicability and has shown that the GDPR applies to deepfakes in many cases because of the wide notion of "personal data." Because "input data" used to create deepfakes generally consists of audio-visual images including faces, the GDPR is applicable to such data.[350] In many cases, a deepfake itself as "output data" also constitutes personal data because individuals featured are often easily recognized. Personal data does not necessarily have to be true.[351]

The result of this is that victims of harmful deepfakes can exercise many different remedies before creators, including the right to have their data erased. However, these remedies are limited in effectiveness because creators are not likely to respect them in the first place and victims might not know who the creators are. Moreover, they are of ex post nature, only being called in by victims once the damage has already been done.

some weblogs and a thesis by Daphne Stevens. See for instance: B.C. Yildirim and C.D. Aydinli, 'Turkey: Deepfake: An Assessment From The Perspective Of Data Protection Rules' (*Mondaq*, 13 November 2019) <https://www.mondaq.com/turkey/privacy-protection/863064/deepfake-an-assessment-from-the-perspective-of-data-protection-rules> last accessed 11 November 2021 and M. Hallé, 'Deep fakes: are there remedies for victims under the GDPR?' (*International Bar Association*, 29 November 2018) <https://www.ibanet.org/Article/NewDetail.aspx?ArticleUid=3ED0EDC2- and D. Stevens, 'Regulating Deepfake Technology' (2020), available at <https://arno.uvt.nl/show.cgi?fid=152071>.

[350] Article 29 Data Protection Working Party, 'Opinion 02/2012 on facial recognition in online and mobile services' (2012) available at <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2012/wp192_en.pdf> p.4.

[351] P. 7 WP 136.

It is for these reasons that a large part of my research in chapter 3 was dedicated to the question whether the GDPR (indirectly) imposes content moderation obligations on social media platforms and search engines regarding the removal of harmful deepfakes. Although I explained that these actors could be regarded "joint controllers" with creators of harmful deepfakes, the "phase-oriented approach" used by the CJEU in the *Fashion ID* case shows us that these actors cannot be held responsible for processing activities where they do not *actually* determine the means and purposes.[352] Thus, the GDPR does not impose content moderation obligations on these actors, because they do not *actually* determine the means and purposes of publications of harmful deepfakes on their platforms. The GDPR's right to be forgotten is the most important mechanism of the GDPR to regulate harmful deepfakes, which should be exercised vis-à-vis social media platforms and search engines. This mechanism is however limited, as it does not prevent harmful deepfakes from appearing online.

In chapter 4, I assessed how the GDPR should be used to regulate harmful deepfakes, in light of other regulatory frameworks. I demonstrated that the GDPR was meant to regulate deepfake technology because it was designed to be technology-neutral[353] and it ensures effective protection of data subjects, in line with the *Google Spain* judgement.[354] To provide effective protection to the rights and interests of natural persons, the main regulatory gap of the GDPR (that it does not provide effective ex ante protection for victims of harmful deepfakes) should be dealt with by applying other regulatory solutions like criminal law and targeted legislation in a complementary way.

In that respect, the GDPR should be used to quickly deal with publications of harmful deepfakes on social media platforms and search engines. Criminal law could be used to identify and prosecute creators of harmful deepfakes in a later stage. Moreover, criminal law in some sense provides ex ante remedies because of deterrence. Using the GDPR and criminal law as complementary instruments however does not prevent harmful deepfakes from appearing in the first place. A solution to this regulatory gap could be to adopt specific legislation that obliges social media platforms and search engines to (further) develop technology that detects harmful deepfakes. However – since deepfake technology will continue to evolve – this solution only works in the short term, and results in a cat and mouse

[352] P. de Hert & G. Bouchagiar, 'Fashion ID and decisively influencing Facebook plugins: a fair approach to single and joint controllership' (2021) Vol. 7(27) Brussels Privacy Hub Working Paper, available at <https://euagenda.eu/publications/fashion-id-and-decisively-influencing-facebook-plugins-a-fair-approach-to-single-and-joint-controllership>, p. 13.

[353] Recital 15 GDPR.

[354] Case C-131/12 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González [2014] ECLI:EU:C:2014:317, para. 60.

game between social media platforms and search engines on the one hand and creators on the other, in the long term.

The answer to the main research question is that the GDPR's right to be forgotten should be used as its main mechanism to offer protection against the harmful use of deepfake technology. Moreover, the GDPR should be a starting point for regulating harmful deepfakes and complemented by other regulatory frameworks such as criminal law and targeted legislation. In this way, the rights and interests of natural persons are most effectively protected and regulatory gaps of the GDPR can be dealt with.

5.4 Implications

My research shows that the GDPR is an excellent starting point for regulating the harmful use of deepfake technology. However, certain problems limit its effectiveness. I have demonstrated that (Dutch) criminal law and targeted legislation could serve as potential solutions for these problems. I have not investigated the potential role of other regulatory solutions.[355] Further substantive research is necessary to assess how problems revolved around the harmful use of deepfake technology can and should be effectively dealt with, especially since existing laws seem unequipped to effectively deal with these issues.[356] Because deepfakes are distributed online, they can easily spread to different jurisdictions. It is therefore evident that effective solutions for dealing with their problems should be provided on an international level.[357]

Another limitation to the findings in this thesis is that it does not cover the publication of harmful deepfakes on the dark web. For this reason, my thesis only deals with the "tip of the iceberg," since it focusses on social media platforms and search engines as actors against

[355] Although I have referred to authors that have done so and have concluded that current legislation is ill-equipped to address deepfake technology. See for instance D. Harris, 'Deepfakes: False pornography is here and the law cannot protect you' (2019) 17 Duke Law & Technology Review 99, where the author explains how there are not sufficient legal remedies in place to effectively prevent harmful pornographic deepfakes and A. P. Gieseke, 'The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography' (2020) 73 Vanderbilt Law Review 1479, p. 1500 or E. Meskys, J. Kalpokiene, P. Jurcys, A. Liaudanskas, 'Regulating Deep-Fakes: Legal and Ethical Considerations' (2019) Vol. 15(1) Journal of Intellectual Property Law & Practice 24, pp. 27-28 where the authors highlight limitations in copyright law and tort law. Other authors argue that existing laws are sufficient to regulate harmful deepfakes. See for instance D. Greene, 'We Don't Need Lew Laws for Faked Videos, We Already Have Them' (2018) <https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them> last accessed 11 November 2021.

[356] Ibid.

[357] On EU-level, the the European Commission has proposed "minimum transparency obligations" for users of AI systems, to disclose that "content has been artificially generated or manipulated when using such AI systems to create deepfakes. See art. 52(3) Artificial Intelligence Act. However, this is not likely to fill the regulatory gaps I have identified.

whom victims can exercise remedies. The dark web is an excellent location for anonymous publication of harmful deepfakes and thus further research is required on this topic.

5.5 Final thoughts

Within the relatively short time span of one year of writing this thesis, deepfakes have become more convincing and have gotten new implementations. A new app called DeepFaceLive now allows users to replace their face with a face of someone else into live webcam footage.[358] Users can use this technology to enter Zoom or Skype meetings, impersonating others. I do not need to explain the potential new problems that may arise. The unavoidable truth is that deepfakes spell trouble.[359] We are warned that more must be done to regulate them.

[358] M. Anderson, 'Real-Time DeepFake Streaming With DeepFaceLive' (2021) <https://www.unite.ai/real-time-deepfake-streaming-with-deepfacelive/> last accessed 11 November 2021.

[359] R. Chesney & D. Citron, 'Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolicits' (2019) 98 Foreign Affairs 147, p. 154.

Bibliography

Articles

Albrecht JP, 'How the GDPR Will Change the World' (2016) 2 European Data Protection Legislation Review 287

Ashworth A and Horder J, *Principles of Criminal Law* (seventh edition, Oxford University Press 2013)

Baxevani T, 'GDPR Overview' (2019), available at <https://www.researchgate.net/publication/333560686_GDPR_Overview>

Bozdag E, 'Bias in algorithmic filtering and personalization' (2013) 15 Ethics and Information Technology 209

Bradford A, 'The Brussels Effect' (2021) 107(1) Northwestern University Law Review 1

Bregler C, Covell M, and Slaney M, 'Video Rewrite: Driving Visual Speech with Audio' (proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, August 1997) <https://dl.acm.org/doi/pdf/10.1145/258734.258880>

Brenner S, 'Distributed Security: A new Model of Law Enforcement' (2005) available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=845085>

Britt MA, Rouet JF, Blaum D, and Millis K, 'A Reasoned Approach to Dealing With Fake News' Vol. 6(1) Policy Insights from the Behavioral and Brain Sciences 94

Carr CT and Hayes RA, 'Social Media: Defining, Developing and Divining' 23(1) Atlantic Journal of Communication 46

Castets-Renard C, 'Algorithmic content moderation on social media in EU law: illusion of perfect enforcement' (2020) 2020(2) Journal of Law, Technology & Policy 283

Chen J, Edwards L, Urquhart L, McAuley D, 'Who Is Responsible for Data Processing in Smart Homes? Reconsidering Joint Controllership and the Household Exemption' [2019] Vol. 10(4) International Data Privacy Law 279

Chesney R and Citron DK, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (2019) 107 California Law Review 1753

Chesney R and Citron DK, 'Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolicits' (2019) 98 Foreign Affairs 147

Chesney R and Citron DK, 'Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?' (2018) <https://www.lawfareblog.com/deepfakes-looming-crisis-national-security-democracy-and-privacy> last accessed 11 November 2021

Citron DK, 'Sexual Privacy' (2019) 128 Yale Law Journal 1870

Cusumano M, Gawer A, and Yoffie D, 'Can Self-Regulation Save Digital Platforms?' (2021) available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3900137>

De Gregorio G, 'The e-Commerce Directive and GDPR: Towards Convergence of Legal Regimes in the Algorithmic Society?' (2019) Robert Schuman Centre for Advanced Studies Research Paper No. RSCAS 2019/36 available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3393557>

De Hert P and Bouchagiar G, 'Fashion ID and decisively influencing Facebook plugins: a fair approach to single and joint controllership' (2021) Vol. 7(27) Brussels Privacy Hub Working Paper, available at <https://euagenda.eu/publications/fashion-id-and-decisively-influencing-facebook-plugins-a-fair-approach-to-single-and-joint-controllership>

Delfino R, 'Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act' (2019), 88 Fordham Law Review 887

Dodge A and Johnstone E, ‘Using Fake Video Technology to Perpetrate Intimate Partner Abuse’ (2018) available at <https://withoutmyconsent.org/blog/2018-04-25-a-new-advisory-helps-domestic-violence-survivors-prevent-and-stop-deepfake-abuse/>
Eichler N, Jongerius S, McMullen G, Naegele O, Steininger L, and K. Wagner, ˊBlockchain, data protection, and the GDPR‘ <https://www.bundesblock.de/wp-content/uploads/2019/01/GDPR_Position_Paper_v1.0.pdf> last accessed 11 November 2021

Eigbedion A, ‘Deepfakes: Legal & Regulatory Considerations in Nigeria’ (2020), available at: <https://ssrn.com/abstract=3670644>

El Zeidy M, 'The Principle of Complementarity: A New Machinery to Implement International Criminal Law' (2002) 23 Michigin Journal of International Law 869

El-Gazzar R and Stendal K, ‘Examining How GDPR Challenges Emerging Technologies’ (2020) Vol. 10 Journal of Information Policy 237

Farish K, ‘Do Deepfakes Pose a Golden Opportunity? Considering Whether English Law Should Adopt California’s Publicity Right in the Age of Deepfake (2020) Vol. 15, No. 1 Journal of Intellectual Property Law & Practice 40

Gieseke AP, ”The New Weapon of Choice”: Law’s Current Inability to Properly Address Deepfake Pornography’ (2020) 73 Vanderbilt Law Review 1479

Goodfellow LJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, and Bengio Y, ‘Generative Adversarial Nets’ (2014) <https://arxiv.org/pdf/1406.2661.pdf> last accessed 11 November 2021

Gorkic P, 'Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW e.V.: More Control, More Data Protection for Website Visitors?' (2019) 5 European Data Protection Law Review 579

Goudsmit MLR, ‘Criminalising Image-based Sexual Abuse: an Analysis of the Dutch Bill against Revenge Pornography’ (2019) 68 Ars Aequi

Gourd E, 'GDPR obstructs cancer research data sharing' (2021) Vol. 22(5) The Lancet Oncology 592

Greene D, 'We Don't Need Lew Laws for Faked Videos, We Already Have Them' (2018) <https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them> last accessed 11 November 2021

Harris D, 'Deepfakes: False pornography is here and the law cannot protect you' (2019) 17 Duke Law & Technology Review 99

Heaton JB, Polson NG, Witte JH, 'Deep Learning for Finance: Deep Portfolios' (2016) Vol. 33(1) Applied Stochastic Models in Business and Industry 3

Karras T, Aila T, Lainen S, and Lehtinen J, 'Progressive Growing of GANs for Improved Quality, Stability, and Variation' (International Conference on Learning Representations, Vancouver, April-May 2018)

Koops EJ, 'The Internet and its Opportunities for Cybercrime' Vol. 1 Transnational Criminology Manual (2011) 735

Koops EJ, Lips M, Prins C, and Schellekens M, 'Should ICT Regulation Be Technology neutral?' (2006) Vol. 9 IT & Law Series, available at <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=918746>

Kugler MB and Pace CL, 'Deepfake Privacy: Attitudes and Regulation' (2021) 116 Northwestern University Law Review, available at: <https://ssrn.com/abstract=3781968>

Leenes R, Palmerini E, Koops EJ, Bertolini A, Salvini P, and Lucivero F, 'Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues' (2017), Vol. 9(1) 1

Maculan E and Gil AG, 'The Rationale and Purposes of Criminal Law and Punishment in Transitional Contexts' (2020) Vol. 40(1) 132

Mahieu R and Van Hoboken J, 'Fashion-ID: Introducing a phase-oriented approach to data protection?' (2019) <https://europeanlawblog.eu/2019/09/30/fashion-id-introducing-a-phase-oriented-approach-to-data-protection/> last accessed 11 November 2021

Mahieu R, Van Hoboken J, and Asghari H, 'Responsibility for Data Protection in a Networked World' (2019) Vol. 10(1) Journal of Intellectual Property, Information Technology and Electronic Commerce Law 84

Masood M, Nawaz M, Malik KM, Javed A, and Irtaza A, 'Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward' (2021) <https://arxiv.org/ftp/arxiv/papers/2103/2103.00484.pdf>

Meskys E, Kalpokiene J, Jurcys P, and Lianudanskas A, 'Regulating Deep-Fakes: Legal and Ethical Considerations' (2019) Vol. 15(1) Journal of Intellectual Property Law & Practice 24

Nguyen TT, Nguyen NM, Nguyen DT, Nguyen DT, Nahavandi S, 'Deep Learning for Deepfakes Creation and Detection: A Survey' (2020) <https://arxiv.org/pdf/1909.11573v2.pdf> last accessed 11 November 2021

Pasetski A, 'Deepfakes: A New Content Category for a Digital Age' (2020) Vol. 29(2) William & Marry Bill of Rights Journal 503

Paun M, 'On the Way to Effective and Complete Protection (?): Some Remarks on Fashion ID' (2020), Vol. 9(1) Journal of European Consumer and Market Law 35

Penney JW, 'Internet surveillance, regulation, and chilling effects online: a comparative case study' (2017) Vol. 6(2) Internet Policy Review 1

Purtova N, 'The law of everything. Broad concept of personal data and future of EU data protection law' (2018) Vol. 10 Law, Innovation and Technology 40

Shih FY and Yuan Y, 'A Comparison Study on Cover-Cover Image Forgery Detection' (2010) The Open Artificial Intelligence Journal 49

Taigman Y, Yang M,. Razanto MA, and Wolf L, 'DeepFace: Closing the Gap to Human-Level Performance in Face Verification' (IEEE Conference on Computer Vision and Pattern Recognition, Colombus, September 2014)

Van Alsenoy, B 'Liability under EU Data Protection Law: From Directive 95/46 to the General Data Protection Directive' (2016), 7 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 271

Van der Hof S,Wraakporno op internet – een verkenning van de (on)mogelijkehden voor een strafrechtelijke aanpak' (2016) 65 Ars Aequi 54

Van der Sloot B, 'Editorial' (2020) 6(4) European Data Protection Law Review 477

Verdoliva L, 'Media Forensics and Deepfakes: an overview' (2020) <https://arxiv.org/pdf/2001.06564.pdf> last accessed 11 November 2021

Vosoughi S, Roy D, Aral S, 'The spread of true and false news online' (2018) 359 Science 1146

Westerlund M, 'The Emergence of Deepfake Technology: A Review' (2019) Vol. 9(11) Technology Innovation Management Review 40

Westling J, 'Are Deep Fakes a Shallow Concern? A Critical Analysis of the Likely Societal Reaction to Deep Fakes' (2019) available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3426174>

Wong R, 'Social Networking: A Conceptual Analysis of a Data Controller' (2009) Vol. 14(5) Communications Law 142

Yazdinejad A, Parizi RM, Srivastava G, and Dehghantanha A, 'Making Sense of Blockchain for AI Deepfakes Technology' (2020) GC Wkshps 2020 1

Zarsky T, 'Incompatible: The GDPR in the Age of Big Data' (2017) Vol. 47, No. 4(2) Seton Hall Law Review 995

Zerlang J, ‘GDPR: a milestone in convergence for cyber-security and compliance’ (2017) Vol. 6 Network Security 8

Books

Brenner SW, *Law in an Era of ‘Smart’ Technology* (Oxford University Press 2007)

Caporusso N, ‘Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology’ in Ahram T (ed.), *Advances in Artificial Intelligences, Software and Systems Engineering* (Springer 2020)

Citron DK, *Hate Crimes in Cyberspace* (Harvard University Press 2014)

Easley D, Kleinberg J, *Networks, Crowds, and Markets: Reasoning about a Highly Connected World* (Cambridge University Press 2010)

Fischer E, *Risk regulation and Administrative Constitutionalism* (Hart Publishing 2007)

Goodwin M, ‘A Dimensions Approach to Technology Regulation’ in Goodwin MEA, Koops EJ & Leenes RE, *Dimensions of technology regulation* (Wolf Legal Publishers 2010)

Leenes R, ‘Regulating New Technologies in Times of Change’ in L. Reins, *Regulating New Technologies in Uncertain Times* (T.M.C. Asser Press 2019)

Moses LB, ‘Regulating in the Face of Sociotechnical Change’ in Brownsword R, Scotford E, and Yeung K (eds), *The Oxford Handbook of Law, Regulation and Technology* (Oxford University Press 2017)

Prosser T, *The Regulatory Enterprise: Government Regulation and Legitimacy* (Oxford University Press 2010)

Case law

CJEU

Case C-101/01 Criminal proceedings against Bodil Lindqvist [2003] ECLI:EU:C:2003:596

Case C-40/17 Fashion ID GmbH & Co. [2019] ECLI:EU:C:2019:629

Case C-40/17 Fashion ID GmbH & Co. KG v Verbraucherzentrale NRWeV (Opinion of Advocate General, 19 December 2018)

Case C-212/13 František Ryneš v. Úřad pro ochranu osobních údajů [2014] ECLI:EU:C:2014:2428

Case C-136/17 *GC et al. v. CNIL* [2019] ECLI:EU:C:2019:773

Case C-131/12 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González [2014] ECLI:EU:C:2014:317

Case C-25/17 Jehovan todistajat [2018] ECLI:EU:C:2018:551

Case C-572/14 Patrick Breyer v. Bundesrepublik Deutschland [2016] ECLI:EU:C:2016:779

Case C-434/16 Peter Nowak v. Data Protection Commissioner [2017] ECLI:EU:C:2017:994

Case C-210/16 Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v. Wirtschaftsakademie Schleswig-Holstein GmbH [2018] ECLI:EU:C:2018:388

Case C-230/14 Weltimmo s.r.o. v. Nemzeti Adatvédelmi és Információszabadság Hatóság [2015] ECLI:EU:C:2015:639

Dutch case law

HR 11 February 2014, ECLI:NL:HR:2014:306

Rb. Gelderland 13 May 2020, ECLI:NL:RBGEL:2020:2521, *NJF* 2020/225

Explanatory reports
Parliamentary Papers II, 2018/19, 35 080, 3

Parliamentary Papers II 1994/95, 23 682, 5

Parliamentary Papers II, 2001/01, 27 745, 3

Parliamentary Papers II, 2000/01, 27 745, 6

EDPS, WP29, and European Commission documents

Article 29 Data Protection Working Party, 'Opinion 02/2012 on facial recognition in online and mobile services' (2012) available at <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2012/wp192_en.pdf>

Article 29 Data Protection Working Party, 'Opinion 4/2007 on the concept of personal data' (2007) available at <https://www.clinicalstudydatarequest.com/Documents/Privacy-European-guidance.pdf>

Article 29 Working Party, 'Annex 2 Proposals for Amendments regarding exemption for personal or household activities' (2013) available at < https://ec.europa.eu/justice/article-29/documentation/other-document/files/2013/20130227_statement_dp_annex2_en.pdf>

Article 29 Working Party, Opinion 10/2006 on the processing of personal data by the Society for Worldwide Interbank Financial Telecommunication (SWIFT)' (2006) available at <https://www.dataprotection.ro/servlet/ViewDocument?id=234>

Commission to the European Parliament and the Council – data protection as a pillar of citizen's empowerment and the EU's approach to the digital transition- two year of application of the General Data Protection Regulation' COM(2020) 364final available at <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0264&from=EN>

European Data Protection Board, 'Guidelines 05/2020 on consent under Regulation 2016/679 Version 1.1 (2020) available at

<https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf
>

European Data Protection Board, 'Guidelines 07/2020 on the concepts of controller and processor in the GDPR Version 1.0' (2020) available at <https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202007_controllerprocessor_en.pdf>

European Data Protection Board, 'Guidelines 3/2018 on the territorial scope of the GDPR (Article 3) Version 2.1' (2019) available at < https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-32018-territorial-scope-gdpr-article-3-version_en>

European Data Protection Board, 'Guidelines 5/2019 on the criteria of the Right to be Forgotten in the search engines cases under the GDPR (part 1) Version 2.0' (2020) available at <https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-52019-criteria-right-be-forgotten-search-engines_en>

Legislation

Council of Europe, 'Convention on Cybercrime' [2003] available at <https://www.europarl.europa.eu/meetdocs/2014_2019/documents/libe/dv/7_conv_budapest_/7_conv_budapest_en.pdf>

Council of Europe, Convention on the Protection of Children against Sexual Exploitation and Sexual Abuse [2007] available at <https://rm.coe.int/1680084822>

Dutch Code of Criminal Procedure

Dutch Criminal Code

Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Proposal for Artificial Intelligence Act) and amending certain Union legislative acts [2021] COM(2021) 206final

Regulation (EU) 2016/679 of the European Parliament and 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 [2016] OJ L119/1

News articles

-- ‘Online filmpjes zijn het nieuwe politieke wapen, maar ‘het effect is beperkt’’ *rtlnieuws* (Hilversum, 25 september 2020) <https://www.rtlnieuws.nl/nieuws/politiek/artikel/5186252/verkiezingen-politiek-campagne-thierry-baudet-pieter-omtzigt-fvd> last accessed 11 November 2021

Holley P, ‘The man who posed as his daughter’s online boyfriend to get nude photos of her’ *The Washington Post* (Washington, 17 March 2016) <https://www.washingtonpost.com/news/true-crime/wp/2016/03/17/the-man-who-posed-as-his-daughters-online-boyfriend-to-get-nude-photos-of-her/> last accessed 11 November 2021

Itzkoff D, ‘How ‘Rogue One’ Brought Back Familiar Faces’ *The New York Times* (New York, 27 December 2016) <https://www.nytimes.com/2016/12/27/movies/how-rogue-one-brought-back-grand-moff-tarkin.html> last accessed 11 November 2021

Kelion L, ‘Deepfake detection tool unveiled by Microsoft’ *BBC News* (London, 1 September 2020) <https://www.bbc.com/news/technology-53984114>

Meyer R, ‘The Grim Conclusions of the Largest-Ever Study of Fake News’ *The Atlantic* (Washington, 8 March 2018) <https://www.theatlantic.com/technology/archive/2018/03/largest-study-ever-fake-news-mit-twitter/555104/> last accessed 11 November 2021

Rothkopf J, ‘Deepfake Technology Enters the Documentary World’ *The New York Times* (New York, 1 July 2020) <https://www.nytimes.com/2020/07/01/movies/deepfakes-documentary-welcome-to-chechnya.html> last accessed 11 November 2021

Online sources

-- ‘Edtech company Udacity uses deepfake tech to create educational videos automatically’ (2019) <https://www.fanaticalfuturist.com/2019/08/edtech-company-udacity-uses-deepfake-tech-to-create-educational-videos-automatically/> last accessed 11 November 2021

-- ‘Article 29 Working Party,’ <https://edpb.europa.eu/about-edpb/more-about-edpb/article-29-working-party_en>, last accessed 11 November 2021

-- ‘How Employers Use Social Media to Screen Applicants’ <https://theundercoverrecruiter.com/infographic-how-recruiters-use-social-media-screen-applicants/> last accessed 11 November 2021

-- ‘Material and Territorial Scope: GDPR Series Part 1’ (2016) <https://www.lexology.com/library/detail.aspx?g=5d778547-bc7e-42b2-acb2-2ec828d40a7d> last accessed 11 November 2021

-- ‘Social Media: Misinformation and Content Moderation Issues for Congress’ (2021) <https://crsreports.congress.gov/product/pdf/R/R46662>

-- ‘This person does not exist: AI generates fake faces on website’ (2019) <https://www.ctvnews.ca/sci-tech/this-person-does-not-exist-ai-generates-fake-faces-on-website-1.4299515> last accessed 11 November 2021

Anderson M, ‘Real-Time DeepFake Streaming With DeepFaceLive’ (2021) <https://www.unite.ai/real-time-deepfake-streaming-with-deepfacelive/> last accessed 11 November 2021

Bond E, ‘AI Video Startup Synthesia Raises USD 12.5m for Multilingual Avatars’ (2021) <https://slator.com/ma-and-funding/ai-video-startup-synthesia-raises-usd-12-5m-for-multilingual-avatars/> last accessed 11 November 2021

Cavalli F, ‘How to detect a deepfake online’ (*Sensity*, 8 February 2021) <https://sensity.ai/how-to-detect-a-deepfake/> last accessed 11 November 2021

Clark B, 'Pornhub promissed to ban 'deepfakes' Videos. And it failed miserably' <https://thenextweb.com/news/pornhub-promised-to-ban-deepfakes-videos-and-it-failed-miserably> last accessed 11 November 2021

Dafoe T, 'Russian Researchers Used AI to Bring the Mona Lisa to Life and it Freaked Everyone Out' (2019) <https://news.artnet.com/art-world/mona-lisa-deepfake-video-1561600> last accessed 11 November 2021

Facebook, 'Data Policy' <https://www.facebook.com/policy.php?ref=pf> last accessed 11 November 2021

Goggin B, 'From porn to 'Game of Thrones': How deepfakes and realistic-looking fake videos hit big' (2019) <https://www.businessinsider.com/deepfakes-explained-the-rise-of-fake-realistic-videos-online-2019-6?IR=T> last accessed 11 November 2021

Greenberg A, 'It's About To Get Even Easier to Hide on the Dark Web' (2017) <https://www.wired.com/2017/01/get-even-easier-hide-dark-web/> last accessed 11 November 2021

Hallé M, 'Deep fakes: are there remedies for victims under the GDPR?' (*International Bar Association*, 29 November 2018) <https://www.ibanet.org/Article/NewDetail.aspx?ArticleUid=3ED0EDC2-92E2-477B-8321-1290AFE00ACC> last accessed 15 April 2021

Hardestry L, 'Explained: Neural networks – Ballyhooed artificial-intelligence technique known as "deep learning" revives 70-year-old idea (2017), <https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414> last accessed 11 November 2021

Heikkilä M, 'Europe's AI rules open door to mass use of facial recognition, critics warn' (2021) <https://www.politico.eu/article/eu-ai-artificial-intelligence-rules-facial-recognition/> last accessed 11 November 2021

Jain A and Lee J, ‘Scars, marks, and tattoos: a soft biometric for identifying suspects and victims’ <https://spie.org/news/1282-scars-marks-and-tattoos-a-soft-biometric-for-identifying-suspects-and-victims?SSO=1> (2009) last accessed 11 November 2021

Jankowicz N, Oits C, ‘Facebook Groups are Destroying America’ (2020) <https://www.wired.com/story/facebook-groups-are-destroying-america/> last accessed 11 November 2021

Knibbe J, ‘Complying with the GDPR on social media – interactions’ (2020) <https://www.linkedin.com/pulse/complying-gdpr-social-media-interactions-jorren-knibbe> last accessed 11 November 2021

Oakes O, ‘’Deepfake’ voice tech used for good in David Beckham malaria campaign’ (2019) <https://www.prweek.com/article/1581457/deepfake-voice-tech-used-good-david-beckham-malaria-campaign> last accessed 11 November 2021

Schwartz O, ‘Deepfakes aren’t a tech problem. They’re a power problem’ (2019) <https://www.theguardian.com/commentisfree/2019/jun/24/deepfakes-facebook-silicon-valley-responsibility> last accessed 11 November 2021

Shakeri S, ‘Lyrebird Helps ALS Ice Bucket Challenge Co-Founder Pat Quinn Get His Voice Back’ (2018) <https://www.huffingtonpost.ca/2018/04/14/lyrebird-helps-als-ice-bucket-challenge-co-founder-pat-quinn-get-his-voice-back_a_23411403/> last accessed 11 November 2021

Yildirim BC and Aydinli CD, ‘Turkey: Deepfake: An Assessment From The Perspective Of Data Protection Rules’ (*Mondaq*, 13 November 2019) <https://www.mondaq.com/turkey/privacy-protection/863064/deepfake-an-assessment-from-the-perspective-of-data-protection-rules> last accessed 11 November 2021

Studies and reports

Adjer H, Patrini G, Cavalli F, and Cullen L, ‘The State of Deepfakes: Landscape, Threats and Impact’ (2019), <https://regmedia.co.uk/2019/10/08/deepfake_report.pdf> last accessed 11 November 2021

De Streel A et al. ‘Online Platforms’ Moderation of Illegal Content Online’ (2020) available at <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf>

European Commission, ‘The GDPR: new opportunities, new obligations’ (2018) <https://ec.europa.eu/info/sites/default/files/data-protection-factsheet-sme-obligations_en.pdf> last accessed 11 November 2021

European Parliament, ‘Tackling deepfakes in European policy’ (2021) <https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf> last accessed 11 November 2021

Paris B and Donovan J, ‘Deepfakes and Cheap Fakes’ (2019) <https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal.pdf> last accessed 11 November 2021