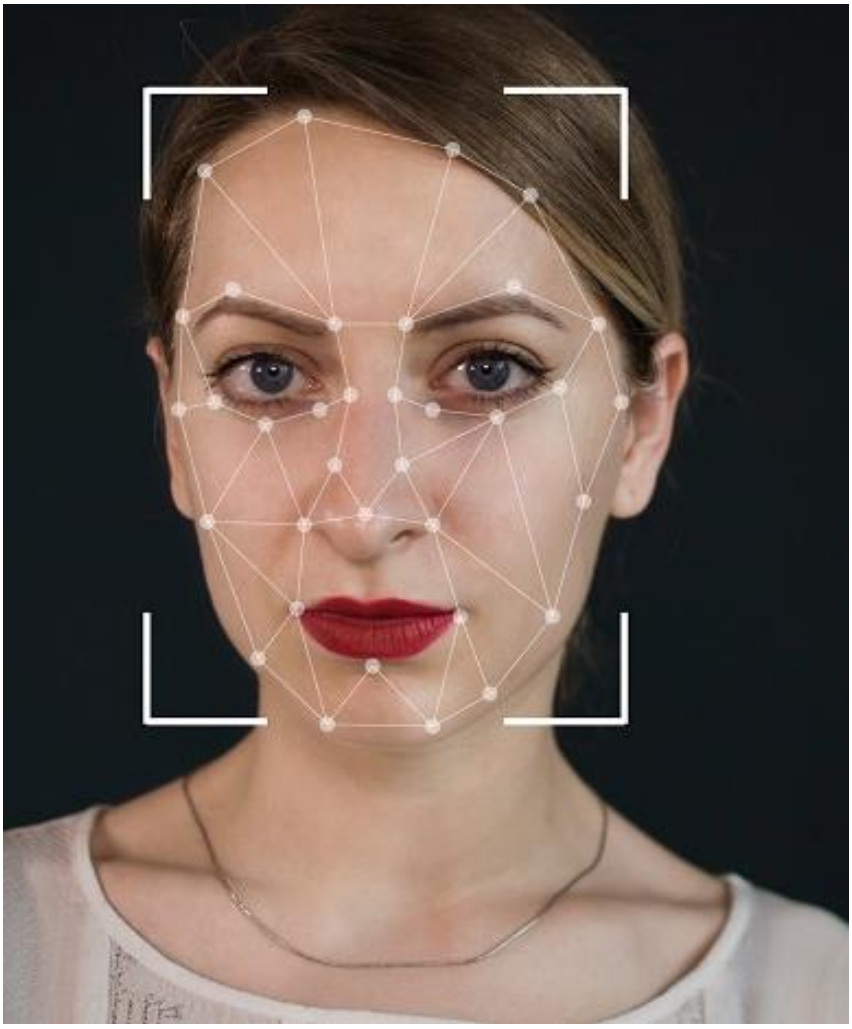
Regulating deepfakes: A legal research on the applicability of regulations on deepfake technology in the Netherlands

Langenberg, Daan

Executive Master Cyber Security

Leiden University



Regulating deepfakes

A legal research on the applicability of regulations on deepfake technology in the Netherlands

Langenberg, D. (Daan)

Thesis supervisors:
Dr. Els de Busser
Dr. James Shires

Preface

Writing a thesis is the final aspect and test of the executive master of Cyber security. As a part-time student, it hasn't always been easy to balance work, a social life, family, and undertaking an academic study. There were periods when the balance was hard to find. It would have been much harder to finish this study without the help and support of my family and friends.

I realise that by writing this thesis, I have only just started doing academic research. This subject will always be interesting to research since it touches on our daily life. It was an insightful and educational experience to write my own thesis, including the ups and downs that come with that.

I want to give a special thanks to my employer to give me the chance to develop myself in the field of cyber security, by giving me the time to work on my study and financing the whole study. I wouldn't have been able to successfully finish this study without the time and financial aid.

And last, but not least, I want to thank my thesis supervisors dr. Els de Busser and dr. James Shires for their guidance, tips, and patience. It probably wasn't always easy to guide a student that is not always fully focused on the thesis, but with your critical and constructive feedback, you have helped me to improve the structure, academic value, and find depth in my thesis subject.

All in all, is this thesis the capstone of a two-and-a-half-year journey that brought me a lot of knowledge, connections, and personal growth. I am keen to use this in the rest of my career and life. I hope that you will enjoy reading this thesis.

Abstract

This thesis is aimed at researching the current regulations on deepfakes in the Netherlands. The goal of this is to identify possible gaps and to provide recommendations for those gaps. This is an important question because deepfakes are relatively new, but there are already some problematic applications of the technology. To tackle those applications, it is important that there is a review of the available and relevant regulations to determine what the best follow-up actions are. The research question of this thesis is as follows:

What legal regulations apply to deepfake technology in the Netherlands and are there any gaps that need to be addressed?

The methodology that this study has chosen is to be an exploratory regulation analysis of the current legal framework. It has done that by reviewing the Dutch Penal code, portrait rights, the GDPR, and regulations on digital platforms that could apply to deepfakes.

The results show that article 139h of the Dutch Penal Code, defamation, portrait rights, and the GDPR apply to problematic applications of deepfake technology. The applicability is concluded by analysing the article and evaluating jurisprudence of the offenses in which photoshop is used instead of deepfakes. So even without having exact jurisprudence on deepfakes on those offenses, it is possible to evaluate the applicability by reviewing the offense in similar cases. Digital platforms are also subject to regulation as they are obligated to delete illegal content as soon as they are aware of it, but they have no obligation to monitor their content. This will change with the Digital Service Act that has been released in 2022 and will be implemented into domestic law in the future.

In the future, the research could focus on the implementation of the DSA, since that will have to balance censorship and human rights such as freedom of speech and access to information. In general, deepfakes are a phenomenon that we will have to keep monitoring to see how their applications develop. There has been some success in the detection of deepfake technology, but this might not be fruitful in the future since the platforms and technology are constantly evolving, and to which the detection software will have to adapt.

Keywords: deepfakes, regulation, Netherlands, creation, publication

Table of contents

Introduction

Deepfakes are an up-and-coming phenomenon that has a lot of different applications. It has gained popularity with impersonations of Obama (Fernandes et al., 2020), but there are also less positive applications like the instant in which the house of representatives in the Netherlands had a conversation with a deepfake impersonation of Navalny, which is the leader of the opposition in Russia (RTL Nieuws, 2021).

Deepfakes are becoming more and more available and are even becoming a part of international conflicts (Argos, 2022; Hancock & Bailenson, 2021). The University College London (UCL) even called it the most serious criminal threat of artificial intelligence (Caldwell et al., 2020). This shows that deepfakes are rapidly becoming part of our reality, but that people are not equipped to determine the authenticity of manipulated media. This is a growing problem since a lot of our information is taken in via the internet and that is exactly the place where deepfakes are being published.

With the growing popularity of this technology, and the fact that it is difficult for the public to identify deepfakes, comes a growing question for regulation. Two parties in the Dutch house of representatives have started working together to initiate a separate law for deepfakes concerning "deepnudes", which are images of sexual nature manipulated with deepfake technology (GroenLinks, 2020).

With every rise of new technology, there comes the question of how to regulate the problematic applications of that technology. This thesis will look at the different applications of deepfake technology, research what regulations are present, and if there are any gaps in these regulations that need to be addressed. By answering these questions, this thesis will help to better understand the landscape of deepfake applications and offer both law enforcement and the public an overview of what regulations apply to certain deepfake applications.

This thesis will continue with the research questions and their relevance. The next chapter will be on the methodology of this research in which also the limitations will be discussed. The third chapter will be focussed on explaining what deepfakes are and how they are created. The fourth chapter will be about the problematic applications of deepfake technology that are identified in this thesis. The fifth chapter will be on the positive applications of deepfake technology to show that there are also useful applications of this technology. The following chapter will be on how deepfakes can be identified, which is necessary if we want to regulate them. The seventh chapter will be on a selection of laws that apply to deepfakes and are therefore seen as regulations on deepfake applications. The eighth chapter will be on regulations that apply to digital platforms on which deepfakes are being published. The final chapter will contain of a conclusion that will discuss and interpret the findings and also contains recommendations for future research.

Research question

The goal of this thesis is to explore the problematic applications of deepfake technology and to determine if the current legal framework of the Netherlands is sufficient to address these applications.

What legal regulations apply to deepfake technology in the Netherlands and are there any gaps that need to be addressed?

To answer these questions, a broad understanding of deepfake technology and different laws is needed. This thesis will answer this research question by examining the undermentioned sub questions.

Sub questions

- What are deepfakes?
- What problematic applications of deepfake technology are included in this thesis?
- Are there positive applications of deepfake technology?
- How can deepfakes be identified?
- What formal laws are present that can be applied to problematic applications of deepfake technology?
- What legal content regulation on digital platforms could regulate problematic applications of deepfake technology?

When all of these questions are answered, it will give a holistic and broad view of the regulations that are present regarding the problematic applications of deepfake technology as discussed in this thesis. From that position, this thesis will be able to determine if the current legal framework has any gaps regarding the specific problematic applications of deepfake technology that this thesis will discuss.

Relevance

Our lives are becoming ever more intertwined with the digital domain. People spend most of their time connected to the online world and it has become an integral part of our existence. This “new” domain is exciting because it is full of possibilities. However, it makes us also vulnerable since we are being confronted with misapplications of these possibilities and this technology. These misapplications are now part of our lives and it is necessary for the public to know what their possibilities are when they are confronted with them.

This thesis is a legal research thesis since it combines different fields of law and includes current content regulation on digital platforms as well. With this holistic approach, this thesis tries to add this new approach to possibly identify gaps or recommendations regarding the struggle against misapplications of deepfake technology. This approach will help research on this subject since it is a new approach compared to other scientific papers regarding this subject.

Deepfake technology has the potential challenges our human tendency to believe that what we see is true. Even in Dutch law, the perception of the judge is one of the means of evidence that can be used in a criminal case. But with deepfake technology, it becomes more difficult to determine whether what we see on video is a factual presentation or if it is altered. The public needs to be equipped to deal with this new technology since misapplication can have dire consequences.

This thesis can also help legal counsel when someone becomes a victim of one of the problematic applications of deepfake technology since it gives insight into the legal measures that are available for the victim. This is important because the digital domain often makes it more difficult for both law enforcement and prosecution. This holistic approach will give them more insight into what their possibilities are.

The last relevant factor of this thesis is that it researches a topic that is still evolving and still at the early stages of its potential. By researching this subject, the phenomenon of deepfakes could receive more attention, resulting in a better understanding of deepfakes and their influence on the general public.

Theory & methodology

Regulation theory of Lawrence Lessig

The different regulations are reviewed in the theoretical framework of Lawrence Lessig's cyber regulation. He states behaviour is regulated by four different kinds of constraints, called modalities. These modalities are laws, social norms, markets, and architecture (Lessig, 1999). Law regulates behaviour because people who don't follow the law will be punished. Social norms regulate behaviour in the way that behaviour will call out a response from others. Norms regulate behaviour because if someone does not comply with norms he will receive punishment by their surroundings. Markets regulate behaviour because they can offer possibilities or restrictions. Prices are influencing the accessibility of products since not everyone will be able to pay for that product. And finally, architecture can regulate behaviour because it can steer people in a certain direction. Examples of this are user interfaces that make it more difficult to make a specific decision and steer people into making the desired decision.

These different modalities work together and will have the largest regulatory effect if different modalities are combined. However, the main focus of this thesis is regulations that fall under the modality of law. This is chosen because laws are the most consistent of the different modalities. Social norms might be different between different groups and cultures. Markets can change rapidly as well, especially in the tech sector. And architecture is very dependent on which platform is used. Of course, laws can change, but it is a long process to change or create a new law. By focussing mainly on the modality of law, makes this thesis relevant for a longer period of time. However, this thesis will also discuss the other modalities to identify the different modalities in the current regulations.

Lawrence Lessig states that law is an order backed by a threat. If someone does not comply with the law, they will receive repercussions and that threat regulates behaviour. Where many believed that cyberspace cannot be regulated because of anonymity and multi-jurisdictionality, Lessig claims that it is possible to regulate behaviour in cyberspace. He claims that there is a direct and indirect effect of the law on behaviour and that architecture becomes the tool of law when the effect of law alone is not sufficient (Lessig, 1999). Regulators will have to balance the four modalities to reach the desired effect because they all affect each other, and will therefore most of the time choose for direct and indirect effects of regulatory modalities.

According to Lessig, the most effective way to regulate behaviour in cyberspace is to regulate architecture, cyberspace itself, or the institutions that create the architecture (Lessig, 1999). That is why his theory is focussed on the effect the law has on architecture and the effect of architecture on law. An example of this is the fact that the law on tapping someone's phone needed to adapt when communication switched to communication over the internet. In that case, the architecture conflicted with the law that was present at that time and needed to be adapted to legally tap someone's communication over the internet (Lessig, 1999).

In his theory, Lessig states that there are three lessons from competition between modalities. The first one is that there is a limit on the power of law to regulate architecture. This depends on who owns the architecture. It is relatively easy to regulate centralized architecture because you only need to regulate one central point. But with software that is "open source" anyone could change the architecture, making it therefore hard to fruitfully regulate the architecture with law. This is also a very large limitation on the power of government to affect architecture in cyberspace.

The second lesson is on transparency. Regulation by law is transparent for the public because the law is widely available to the public. But it is not always transparent in what way law regulates the architecture of software. The user is, without its knowledge, indirectly being regulated by law, because the law regulates the architecture.

The final lesson is that the regulation needs to be specifically tailored to its goal. Regulating architecture changes the foundation of software and should therefore the regulation should be precisely tailored to avert unwanted consequences.

Based on this theory, I expect the presence of several regulations that would affect deepfakes in the Netherlands. I expect that, by the theory of Lessig, the biggest effect of regulations can be created by law and architecture and that the regulations applicable to deepfake technology will be mainly of these two modalities. I expect that the modality of law influences behaviour directly an that it will affect behaviour indirectly as well through architecture. I assume that cyberspace is not well regulated since there is no jurisdiction of governments, but that it will be mainly covered by architecture. This means that I expect to find mainly regulations concerning architecture. And if there are laws that directly regulate behaviour, I would expect them to be general in use and not specific to cyberspace.

Methodology

The objective of this research is to evaluate the current regulations concerning deepfake technology in a holistic approach. Holistic research differentiates itself from different forms of research in a way that it opens up the possibility to look at a phenomenon from various different perspectives. This method of research has been used in various research and offers the researcher a wider perspective (Oktay & Bala, 2015; Woodley & Lockard, 2016; F. Zhao et al., 2021). As this research method allows me to pay attention to more regulations than solely criminal law, it could provide a more comprehensive perspective and could therefore help identify if there are lingering gaps. This means that it will include an array of legal frameworks, societal research, and content regulation regarding different problematic applications of deepfake technology. However, there is a risk with trying to generate a broader perspective. A broader perspective could end up in a lack of depth in each of the different components. To tackle that risk, this thesis will select a few specific regulations from different fields to ensure that there is a holistic approach with enough depth for the different components.

The nature of this research is to be an exploratory regulation analysis. It aims to explore what regulation is present and applicable to deepfake technology in the Netherlands at this time. This research is done in an inductive method. It takes the current legal framework and regulation as a starting point to determine its applicability to different appearances of deepfake technology.

The main sources for this research will consist of academic literature done in the technical, legal, and societal domains. Policy texts, formal law, and jurisprudence will also be included to assure a broader take on the subject. Formal law is considered to be law that is established by an institution according to certain processes. This is the opposite of informal law, which involves unwritten customary regulations which can be present (Friedman & Hayden, 2017). Not all formal laws will be evaluated in this matter. As I will explain in more detail later on, laws that are examined in this thesis are chosen because they will offer a broad perspective on regulations that are applicable to deepfakes. These texts will be analysed to determine what regulation on deepfakes is present in the Netherlands and if that is sufficient for the applications that we see today.

Before this determination can be made, it is important to set a working definition and description of what are considered to be deepfakes. This is done by examining the deep learning technology that makes deepfakes technologically possible. This technology can be used for both rightful and problematic applications and this research will examine which applications are present at this time.

After the evaluation of the different applications of deepfake technology, this thesis will examine the different regulations that are present and evaluate if they are applicable to those applications of deepfake technology. These regulations consist of laws and content regulations on digital platforms. These laws and regulations don’t have to mention deepfake or deepfake technology specifically. The evaluation will be focussed on regulations in which the application of deepfake technology could be applicable. This thesis will also include, for each regulation, why the definition of deepfakes is applicable. This is a thin line with a large grey area, but boundaries need to be set to make sure that the research is feasible. Therefore, after careful consideration, a selection of regulations has been made based on the following two criteria:

- Regulations regarding the use of images;
- Regulations regarding the problematic uses of deepfake technology described in this thesis;

The first criterion includes regulations from criminal, portrait law, the General Data Protection Regulation (GDPR), and content regulation on social media and is primarily focussed on offenses in which technology is used to create, alter, or share imagery. The second criterion is broader since it is aimed at the most common problematic uses of deepfake technology. In practice this allows the thesis to include defamation, which is the most commonly used regulation with regard to the problematic uses of deepfake technology included in this thesis. The other laws are already selected under the first criterion. This will lead to a broader view of the regulation that is present and will give the possibility to determine where there are possibilities for future research and regulations.

This selection is made because it offers an approach that combines several aspects. It is a combination of national and international regulations, where the penal code and portrait rights are national and the GDPR and deepfake content regulations on digital platforms are international. The selection of regulations is also complementary to each other because it is a combination of formal laws in the Netherlands and different regulations. The penal code and portrait rights are established by the government and are actually a law that can be prosecuted. The GDPR and the deepfake content regulation on digital platforms are different forms of regulation since they are not directly implemented in the Dutch legal framework. These regulations are chosen because they will offer a perspective from these different regulations. There are more regulations that could be applicable to deepfakes in the Netherlands, but this specific selection is chosen because it will ensure it to be a comprehensive legal research with enough diversity in the different regulations to offer a broader perspective.

This method has been chosen because of several reasons. Deepfake technology is still evolving and is a relatively new phenomenon. There is some scientific literature on the subject, but a lot of them focuses on a very specific area, for instance, revenge porn. Since it can take a lot of time to adapt existing legal frameworks and regulations to a new phenomenon, it is important to create a holistic view of the existing research to identify existing gaps and opportunities for future research and regulations.

I have chosen to include both formal laws and online content regulation because both show promise in their applicability to deepfake applications. The legal research will limit itself to existing laws in which deepfake applications are a core part of the offense. This means that the offenses are committed by creating, altering, or sharing media with deepfake technology. This means that it will include portrait rights, but it will exclude criminal offenses like blackmail. Blackmail is a criminal offense in which deepfake technology is just one of the methods and therefore it is not interesting to include these criminal offenses because it can be argued that the Dutch Penal Code is applicable in those cases. The same goes for criminal offenses like extortion since the threat of creating or spreading deepfake material is not criminalized, but the fact that it is used to get something from someone is a criminal offense.

This is, however, still mostly theoretical since there is, at this moment, no jurisprudence in which it is determined that these regulations are applicable to deepfake technology. These regulations are so-called technology-agnostic. This means that these regulations have no fundamental connection to technology but that it is a regulation with broad applications. This is contrary to regulations like the criminal offense of hacking which is described in article 138ab of the Dutch Penal Code. In that criminal offense, technology is a fundamental part of the criminal offense. But there are no specific regulations in the Dutch Penal Code yet that are focussed on deepfakes. Therefore, the focus will be on technology-agnostic regulations which are present at this time.

The main objective of this research is to gain insight and to give a holistic overview of the existing regulations regarding deepfakes. This would give insight into, if any, gaps in the current legal framework or possibilities for existing regulations. Since the regulatory framework keeps changing and deepfakes keep evolving, it is important to keep reviewing the different regulations that apply to these or new problematic applications of deepfake technology. Regulations that are present and effective today, might be overruled in the future.

Limitations

Lack of jurisprudence

A limitation of this research is that it is mostly theoretical when it comes to formal laws because there is no jurisprudence yet in which deepfakes are part of the court case. This is a limitation because the law is made by politics, but then it is formed by jurisprudential texts. Therefore, the evaluation of the applicability of laws to the applications of deepfake technology is purely based on prior research and my own analysis. I will look at the whole scope of the law to be as thorough as possible. This means that the sources will consist of main texts, critical law reviews, and jurisprudence of similar cases. Future research will be necessary when there is jurisprudence on this subject since that might give new insights on how to tackle the problematic applications. The jurisprudence that this thesis will include focuses on cases that are comparable with the results of deepfake technology, like photoshop.

Geographical focus

This thesis is only focussed on the legal framework that is applicable in the Netherlands. For future research, it might be interesting to analyse how different nations are handling the problematic applications of deepfakes as it might be transferrable to the situation in the Netherlands.

Early in regulatory developments

There has been a surge in regulations concerning online content compared to the early stages of the digital age. This means that this thesis is being written at a stage in which new regulations can appear suddenly giving new insights for future research regarding this subject.

Personal bias

Due to my background as a military police officer, I am inclined to trust our legal system. I also have the tendency to view formal laws from the perspective of law enforcement. This bias might influence this thesis in the recommendations and discussion section.

Second, I have little experience in the field of academic research and am still learning on this matter. Thankfully I am being guided by two supervisors in my research, helping me forward and tackle my lack of experience.

Selection of regulations

This thesis has made a selection of what regulations are included in the research. These regulations aren’t the only regulations that are present regarding deepfakes. There had to be a selection to make this thesis feasible, but it might give some interesting insights to research different regulations as well. Looking at regulations that aren’t included in this thesis might complement this thesis.

Nature of deepfakes

Deepfakes refer to all different kinds of synthetic media where the person is swapped with the characteristics of another person. The name "deepfake" first emerged on Reddit in 2017 when an anonymous user created it by combining the terms "deep learning" and "fakes" (J. Kietzmann et al., 2020; Maddocks, 2020; Somers, 2020). This is mostly done by using software that uses a technology called "deep learning" or "machine learning" and artificial intelligence. In the beginning, most targets for deepfakes were famous people like actors and politicians and it was mostly used as a gimmick to make politicians give different speeches like the example in 2017 called "Synthesizing Obama"(Suwajanakorn et al., 2017).

The term deepfake is often used as an umbrella term of manipulated media, so it would include altered images with photoshop. This thesis focusses specifically on media that has been created or altered with deepfake technology. This means that media altered by photoshop, or misattributed media is not considered to be a deepfake in this thesis.

Since then the technology that is needed to make deepfakes has developed and nowadays it is quite easy to make a deepfake video via apps like "Wombo", "FaceApp", "DeepFaceLab", and "Deepfakes web β". This software is becoming more and more successful because it is easily available and the quality is increasing at such a rate that the results are becoming ever more believable.

This believability is an aspect that has a significant impact because we, as humans, are placing a lot of trust in the things that we see and hear (J. Kietzmann et al., 2020). Of course, people are now more aware of the fact that visual media can be compromised, altered, or fabricated, but according to prior research, it is still one of the means of evidence that we trust the most (Granot et al., 2018; Porter & Kennedy, 2012). Meaning that this technology can influence our decision-making since it can alter our perception of what is authentic. When we compare early deepfake material, like the example in which the face of Steve Buscemi was placed on the face of Sharon Stone in the movie Basic Instinct, and a newer example of Ross Marquand on Jimmy Kimmel Live, we can see how far the technology has developed and that is becoming ever more difficult to distinguish what is real and what is altered.

This combination of the technology becoming more easily available and the results becoming even more convincing is fuelling its popularity and is also making the case on why it is important to look at this phenomenon from both the benefits it can bring and the problematic applications of this technology.

Deepfakes are created by using technology that is based on machine learning and artificial intelligence (J. Kietzmann et al., 2020; Maddocks, 2020; Somers, 2020). This technology can identify facial movements in the original media and can transfer those movements to a different face, making it possible to make it look like that person is the one in the video. It can even use this technology to learn the voice of both persons and create a soundtrack in which that other person is saying the same thing in their own voice. The next section will give a short explanation of how this software is used to replace the face of someone in a video to create a convincing deepfake. Not by simply using an application that does all the work for you, but by looking at the technology called deep learning that makes deepfakes possible.

This technology can have both beneficial and troubling applications in real life which are discussed later in this chapter. It is important to keep in mind that the nature and application of this technology are still evolving and could be entirely different in a few years. That is why this technology must stay a subject for debate so future developments are continuously discussed to determine whether or not the current regulations on deepfakes are sufficient.

Deep learning and deepfakes

The basic of deepfakes is a technology called deep learning. Deep learning is a variant of machine learning and is focused on a form of artificial intelligence that can train a computer comparable to how we train a human brain. Deep learning uses a deep neural network (DNN) which consists of a large compilation of artificial neurons and each of these neurons performs a simple computation. But by working together they can solve more complex problems like recognizing facial expressions or specific persons on a screen (J. Kietzmann et al., 2020; T. C. Kietzmann et al., 2019).

In the human brain, the connections between neurons are vital for the learning process. It strengthens connections between neurons that are needed, increasing the efficiency in which the neurons are cooperating and eventually improving our performance in the task that is related to that connection (J. Kietzmann et al., 2020). An example of this is that at first, you might find it difficult or near impossible to solve a Rubik’s cube, but with intense practice, you can solve the puzzle with a lot less effort because your brain has learned how to solve it. The DNN is working in a similar way that the connections between the neurons need to be trained to improve the operation of that system. When an untrained DNN is given 1000 pictures of facial expressions, it would not be able to identify which picture is related to which facial expression. A trained DNN has better-trained connections between the neurons and will be able to identify underlying characteristics of a face and it will be able to recognize facial expressions (Goodfellow et al., 2016; J. Kietzmann et al., 2020).

DNNs can be trained by giving the DNN a vast amount of data consisting of images with the correct facial expression. It can identify which underlying characteristics are specific to that facial expression so that when the DNN is encountered with a new photo, it should be able to identify which facial expression is displayed. And every time that the DNN links the wrong facial expression to an image, it learns from that mistake, making the DNN increasingly accurate. This process of deep learning needs a vast compilation of videos and images, i.e. a training set, making it quite an extensive process (J. Kietzmann et al., 2020). It is also interesting to look at the privacy of the people that are used in the training sets. However, that is focussed more on harm inflicted by the creation of the algorithms of deepfakes instead of harm inflicted by deepfakes. Therefore it will not be included in this thesis. When the process of deep learning is executed well, it results in a system that gives the impression that it has intelligence, hence the term artificial intelligence.

One important aspect of this process is something called the “autoencoder”. The autoencoder is the part that is trained to recognize aspects of an image, like faces, based on a large database of images and videos. The autoencoder can start to use the recognized aspects to create a new image based on characteristics that it has learned from all the other images (Goodfellow et al., 2016; J. Kietzmann et al., 2020). This starts by first encoding the separate aspects of an image, like the position of the eyes, mouth, skin colour, etc., into different measurements. These measurements can be set in advance and are dependent on how many aspects are desired by the programmer. It does this because all the data that needs to be compressed to the DNN can learn to reconstruct images from fewer data than the original image. This compression continues until there is a so-called information bottleneck, or latent space, which contains the necessary measurements the DNN needs to reconstruct the image. After this, there is the decoder process which has the purpose of creating an image from the information bottleneck that resembles the original image.

The autoencoder is trained to compress original images in different measurements and to recreate images with different facial expressions. But with only this process it is not possible to instruct the program to create an image with a smiling person because in this process it is not predictable what combination the technology will make. To do this, we will need to combine the latent space of different images to create a shared latent space. This can be done by having two images of people used on the same encoder, but then decoded by a person-specific decoder. With this method, the encoder will learn to use features that both images have in common, like facial expressions. This will result in similar measurements in the latent space, making it possible to reconstruct the face and facial expression of another person on a different image. That outcome will generate an image that is fully constructed, but the facial expression and other details will be the same as the original image (J. Kietzmann et al., 2020).

Applications of deepfakes

Problematic applications of deepfakes

As with all new technologies, it can have both positive and negative applications. This is the same with the technology of deepfakes and this paragraph will discuss the problematic applications of this technology. The next paragraph will discuss the positive applications of deepfake technology.

The first problematic application of deepfakes is that they could be used for crimes involving identity theft (Citron & Chesney, 2019). With deepfake technology, people could impersonate others in the digital world and they could use that identity for financial or a different kind of gain. This could have a variety of consequences in the physical world because our digital lives are entangled with our physical lives.

A large market for deepfake technology is pornography. Deepfakes have been used from the beginning to place the faces of others on pornographic material for several reasons. It could be that the creator has the intention to use the material for his or her own pleasure, but it is also very prominent in revenge porn which has the purpose of inflicting harm to the person that is portrayed (Citron & Chesney, 2019; Maddocks, 2020). It is possible to manufacture pornographic material of people based on pictures that are available via social media like Instagram, making it quite easy to make this material without the consent of the victim. The effects that revenge porn has on victims are worrying and could lead to PTSD, anxiety, depression, suicidal thoughts, and other mental health issues (Bates, 2017). The majority of victims of cyberstalking and non-consensual pornography are women (Eaton et al., 2017), so it is in line with expectations that the majority of victims of deepfakes in pornography are female as well. There are already some examples of women who were victims of deepfake pornography because they were speaking out against a social issue. These were created to silence those women and can be seen as a threat to the emancipation process of women because it denounces them as objects to silence their arguments (Kerner & Risse, 2021).

Deepfake technology is also a very effective means for sabotage. Deepfakes can be used to sabotage rival companies by displaying them in a shocking video. An example of this could be that employees of a certain company could be shown in a private company in which they are using racial slurs or other means of discrimination. In today's world, the image of a company is very important. In the past, we have seen that companies have suffered real consequences when material like this appears because people will start to boycott your company in a movement which is called "cancel culture" (Ng, 2020). This can happen in a very short time frame, making it impossible for those companies to expose the material as fake. And even if the material was later exposed as fake, it is not certain that everyone will be informed that the material was fake and the company could have long-term consequences for its image.

Deepfakes can also be used to sabotage your personal rivals at work. A lot of companies screen the social media of the applicants when they are hiring. Making a deepfake video of your rival and placing that on the internet, might result in the fact that your rival will not be invited for a job interview because his or her profile was discarded in the screening based on the deepfake material (Alexander et al., 2019). This application of deepfakes can have consequences in the long term because the victim will have to prove that the material is fake, which might be extremely hard to achieve.

A different application of deepfakes is that they can be used to influence and disrupt public discourse. There is a need for shared facts and truths for public debates to have substantive contributions. If there is a lack of those shared facts and truths, most debates will just be monologues of each other's point of view and will most likely result in a polarisation of society. With deepfakes, the whole foundation of shared truths and beliefs that is necessary for a substantial public debate is questionable because it is harder to establish what is true (Citron & Chesney, 2019). If there is video evidence, people might assume that it is real, but with this technology, it could be manufactured making the debate shift from content to the value and validity of evidence and that could fuel polarisation within society. Deepfakes can also be used to create evidence for conspiracy theories and allow people to live in their own reality where their beliefs are being validated by manufactured evidence.

As we have seen in the examples above, a large issue with deepfakes is that they can convince people something is true based on manufactured evidence creating a link with misinformation campaigns. There are several instances in recent history where state actors tried to influence the elections in another state by spreading misinformation to its population (Westerlund, 2019). Misinformation is making the work of journalists more difficult, as we have seen after the attack in Christchurch where there were videos shared that were supposed to be footage of the shooting, but it was later discovered that it was footage of a different incident. Journalists have incentives to report instantly after an event because it can give them an edge over their competitors. But this conflicts with their other incentive to report truthfully. With deepfakes, it can be harder to distinguish what is real and what is fake so journalism is facing a delicate dilemma when they encounter evidence of an event (Westerlund, 2019).

And finally, deepfakes can be used to hamper the trust that citizens have in official institutions. Employees could be portrayed in a deepfake video in which they are performing a shocking act, like discrimination or human rights violations. This is also the case for politicians where deepfakes could severely damage their popularity during the elections. This material could not only lower the trust that the general public has in official institutions but it could also fuel demonstrations and uprisings (Citron & Chesney, 2019; Westerlund, 2019).

These are just a few examples and they should not be seen as an exhaustive overview of the problematic applications of deepfakes. These examples have been chosen because they might be susceptive to the regulations that are included in this thesis. The technology is still evolving and there will be different applications, which are not mentioned in this chapter, that are significant enough that they should receive attention as well. Therefore, it is important to continue reviewing the applications of deepfakes and we should keep questioning ourselves if those applications are something that we want to promote or impede.

Positive applications of deepfakes

Besides the problematic applications of deepfake technology described in the previous paragraph, it does offer very positive applications as well. The first application of deepfake technology is its usage in mourning therapy (de Ruiter, 2021). The idea of using this technology in mourning therapy is that the relatives can see the deceased again to help them process their mourning. People might have certain things on their minds that they weren’t able to say when the deceased person was still alive. The Dutch public broadcasting company KRO-NCRV made a documentary in 2020 in which they experimented with this application of deepfake technology (KRO-NCRV, 2020). In cooperation with a mourning therapist, relatives can have conversations with deceased loved ones. It shows very touching moments of relatives having conversations with their deceased loved ones and it shows how convincing this technology can be resulting in very special experiences for the participants.

Using deepfake technology to “bring deceased people back to life” is one application of this technology that has a large ethical component. In this case, it is created with the consent of the relatives and used in their mourning process. The company MyHeritage has been offering it as a service since 2021 called "Deep nostalgia" and they use deepfake technology to animate old family photos, making it more of a gimmick. The deceased ones cannot give consent to how and if their image is used in this technology, making it an ethical dilemma how far the relatives can go with this technology.

Positive applications of deepfake technology also include helping victims of diseases or injuries cope with those disabilities (J. Kietzmann et al., 2020). An example of this is that a victim of the disease ALS could lose their voice and with deepfake technology, people could get their voice back (de Ruiter, 2021). A different, more controversial application, is to use deepfake technology in porn for people who aren't able to have certain experiences themselves. People might have disabilities that limit their possibilities to enact in certain scenarios and this technology could give them, even though it is virtual, experiences that otherwise would never be possible (Volpe, 2018).

Deepfake technology can also be a game-changer in the field of education. Nowadays most forms of education consist of a large part of lectures and reading. Some students have more trouble with this form of education and deepfakes could increase the interactivity of teaching material (Citron & Chesney, 2019). An example of this is that historical figures could be brought back to life and students could have actual conversations with them, providing the students with the same teaching material but in a manner that might be more efficient for some students. Prior research has shown that using applications of virtual reality can be beneficial for the student from a pedagogical perspective (Citron & Chesney, 2019).

And finally, there are commercial applications of deepfake technology that could be used like making it possible to try out different hairstyles at the hairdresser, trying out glasses from home instead of going to the optician, and trying different kinds of cosmetics without having to apply everything. Deepfake technology can also be used in movies to display the face of the real actor on the stunt double to make the stunts even more convincing (J. Kietzmann et al., 2020; Westerlund, 2019).

These positive applications of deepfake technology are making it clear that simply banning all the regulations, even if that would be possible, wouldn't be a desirable outcome. That would mean that the potential of these positive applications are overlooked because there are also problematic applications. This shows that, ideally, there should be enough regulation to tackle the problematic applications of deepfakes, while leaving enough room for the positive applications to flourish.

It is important to note here that the technology of deepfakes is still evolving and that there are a lot more positive applications of deepfake technology that aren't listed in this paragraph. But the purpose of this paragraph was to show that there are very useful and positive applications of deepfake technology that we can embrace and instead of letting the problematic applications of this software control the debate. We should keep examining the positive applications of this software and find a balance in the regulations to mitigate the harm of problematic applications while maintaining the positive applications.

Typology

These different applications have differences and similarities when they are evaluated from a goal perspective. The problematic applications can be divided into two categories. The goal of the problematic applications in the first category is to gain a personal advantage. The second category is focussed on inflicting harm. The purpose of this typology is to identify the differences and similarities and to put them in perspective.

The first category includes identity theft and sabotaging others or different companies. With these problematic applications, one could gain an illegal advantage on someone else or at the expense of someone else. With identity theft, people could gain a variety of advantages, from transferring money to gaining access to certain information. Sabotaging others could give someone an advantage because the reputation of others, or other companies, might be infringed so your position improves in the field of competition.

The second category includes revenge porn and hampering trust in official institutions. Here deepfakes are used with the goal to inflict harm on others or organisations. Revenge porn is often used to shame others and therefore inflict harm on that person. And creating deepfakes to hamper the trust people have in official institutions is purely focussed on harming the reputation of that institution.

The problematic applications of influencing and disrupting public discourse can fit in both categories. As we have seen with Alex Jones, it is possible to monetize polarisation in society and deepfake technology could be used to create evidence of doubtful conspiracy theories. Alex Jones is the owner and presenter of the website InfoWars which openly spreads conspiracy theories. He has been sued by the families of victims of the Sandy Hook Massacre because he has been openly stating that the massacre didn’t happen and that it was a conspiracy (The New York Times, 2018). He has recently been found guilty in a defamation trial and is ordered to pay 49.3 million dollars. This is based on the amount of revenue he made from making those claims. Hypothetically, these theories could be made more believable by using deepfakes and could therefore result in more attention and exposure to the public which increases possible revenue from advertisements or merchandise.

Gaining advantage	Inflicting harm
Identity theft	Revenge porn
Sabotaging others or organisations	Damaging trust in official institutions
Influencing and disrupting public discourse	

Table 1: overview categories problematic applications of deepfake technology

The positive applications can also be divided into two categories based on their goal. Mourning therapy and dealing with disabilities can be included in the category of processing harm. The second category is the category of entertainment which includes the application for education and trying on new hairstyles. With mourning therapy and coping with disabilities, deepfakes can play an important role as described in this chapter. This is a very noble and positive application of deepfake technology because it can help people to cope with their loss or harm. The second category of entertainment has the goal to entertain people. Both applications are fun and useful applications that could make life better and more enjoyable which offers a market for organisations to sell deepfake technology for monetary profits (Zannettou et al., 2019).

Processing harm	Entertainment
Mourning therapy	More engaging education
Coping with disabilities	Trying out hairstyles

Table 2: overview categories positive applications of deepfake technology

It is clear that there is a need for regulation for problematic applications. However, the positive applications of deepfake technology are making it clear that the technology can have a positive purpose in life and should therefore be facilitated where possible. Regulations should therefore find a balance in which the problematic applications are mitigated and the positive applications can be utilized.

How can deepfakes be identified?

To tackle the problematic applications of deepfakes, it is detrimental that those deepfakes can be detected. Methods for detection have been proposed from the moment deepfakes started emerging and the first methods were very reliant on individuals while newer techniques rely on the same technology with which deepfakes are made, i.e. deep learning, to automatically extract unique features to determine if the content has been altered (Nguyen et al., 2021; Westerlund, 2019).

The problem with deepfake detection solutions is that makers of deepfakes are using those solutions to make their deepfakes even more convincing. Aspects of the technology that can be detected are becoming increasingly difficult to spot because the developers of deepfakes are trying to make their software give the most convincing output, making it a cat-and-mouse game. Additionally, there is more effort in developing deepfake technology than detection technology (Westerlund, 2019).

The largest contribution could come from technological solutions and there have been studies concerning several different approaches to detecting deepfakes. But most studies conclude that the detection methods are still having a large error rate and are unable to detect deepfakes (Nguyen et al., 2021). There are however models based on deep learning called Generative Adversarial Networks (GANs) that show promising results in detecting deepfake images (Europol, 2022; Hsu et al., 2020). With videos, it is more difficult to use detection methods because there is a loss of frame data when a video gets compressed (Nguyen et al., 2021).

Detecting deepfake videos is also more difficult because it is more complex than a single image. In an image, there is only one frame that can be investigated and with a video, there are substantially more frames that have to be investigated. But it is not impossible and an example of a method that can be used is the ABC metric (Fernandes et al., 2020). This method doesn’t require access to the training data and it doesn’t need to be trained on separate validation data, which makes the method simpler to use on different content. On three different datasets, the ABC metric reached a 96% accuracy, making it fairly accurate at detecting deepfakes. And there are several other researchers that claim to have created deepfake detection methods that show around the 90% or more accuracy (Guarnera et al., 2020; Guera & Delp, 2019; Tariq et al., 2021; T. Zhao et al., 2020) but all of them end in the discussion with the notion that their method needs more extensive testing before it can be generalized.

Besides looking at the images of a video, there are also developments in detecting deepfakes based on audio as well. When we communicate, a lot of emotion is shared verbally and that data can also be analysed to see if the content is manufactured or manipulated. This perspective has been included in research that checked both the video and the audio to determine if the content was a deepfake. They tested their deep learning network and found that in the first test they had an 84.4% level of accuracy and in the second test they achieved a 96.6% level of accuracy (Mittal et al., 2020).

In summary, it is possible to detect deepfakes using the same technology which is used to create deepfakes, i.e. deep learning. It is however still evolving and not yet available for the general public, while the software to create deepfakes is publicly available. This shows that the development of detection methods and technology lags behind the development of deepfake creation technology. The detection methods do have promising statistics, but even if the technology becomes so accurate that almost 100% of the deepfakes can be detected, it then still needs to be implemented in all our devices and services that we use on our digital devices. These are still two giant barriers that need to be addressed if we ever want detection methods and technology to be widely available to the public.

Laws regulating deepfakes

Discussing the problematic applications of deepfake technology makes it apparent that there is a need for some kind of regulation. This chapter will evaluate what existing laws are available for countering the problematic applications of deepfakes. The chapter starts with an examination of article 139h of the Dutch Penal Code, which came into effect in 2020 and was made to counter the rising phenomenon of revenge porn. Next, this thesis will examine the criminal offense of defamation and its applicability to deepfakes. Subsequently, this thesis will make the switch to civil law and evaluate portrait rights and the GDPR and how this is applicable to deepfakes. Then, this chapter will analyse the similarities and differences between these regulations and identify common denominators. This chapter will conclude with a conclusion.

Article 139h Dutch Penal Code

Creating pornographic material of someone without their consent is, since January 2020, a criminal offense under article 139h of the Dutch Penal Code (Openbaar Ministerie, 2020). That article says that whoever deliberately and illegally creates an image of sexual nature of someone, or whoever has the access to such an image whilst knowing, or has reason to suspect that it was illegally obtained will be punished with a fine of the fourth category, which is a maximum of €22,500 (Rijksoverheid, 2022), or detention of one year. The key word here is “illegally” and in this regard, this implies that the person who is the subject of that material didn’t give consent for the creation or distribution. The sentence can even be increased to the detention of two years if the content is disclosed to others whilst knowing that it has been obtained with a criminal offense or if you disclose it to others whilst knowing that disclosing the image might harm the portrayed person.

When this article was formed, not everyone was immediately convinced that there was a need for a specific criminal provision for “revenge porn” (Berndsen, 2020). There were already possibilities like defamation and insult, which is a criminal offense in the Netherlands. However, the government was of the opinion that the existing criminal provisions didn’t properly cover the offense. Having a specific criminal provision for “revenge porn” would contribute to an unambiguous criminal law approach and could also ensure recognition of the suffering inflicted on the victims. As discussed earlier, this kind of harm can have extremely long-term effects on the psychological condition of the victim (Bates, 2017).

Also, by creating a specific criminal provision, the government states that it condemns this kind of behaviour (Berndsen, 2020).

If we look at this criminal offense it is instantly clear how deepfakes can fit this description among many other forms of revenge porn. Creating an image of sexual nature of someone using deepfake technology can be considered to be illegally creating an image of sexual nature. The illegal part focuses primarily on the fact that the person who is portrayed does not give consent. This is both the case in “revenge porn” where someone spreads intimate images of his or her ex and in the case of creating deepfakes to make them identifiable in pornographic material.

The interesting aspect of this criminal offense is that there is a difference between creation and publication. There is a maximum of one year detention if someone illegally creates or has access to images of sexual nature. But this sentence can be doubled to two years if someone discloses that material to others. This demonstrates that the legislator evaluates the seriousness of the criminal offense in accordance with the effects. There is still harm if the content is created of someone, but that harm is significantly worse if that content is also shared with others. There is however an important notion here that there is no intent needed for sharing in this criminal offense. Even if someone accidentally shares the content with someone else, it is still considered to be sharing and that person can then be prosecuted for that.

This new article in the Dutch Penal Code gives the public prosecution new options (Tweede Kamer der Staten-Generaal, 2019). Before this article, most “revenge porn” cases were dependant on reports of the victims and most cases had to be prosecuted for libel and slander. However, it was very difficult to prosecute for libel and slander because there is a specific need for intent in inflicting harm to the victim. Therefore, a lot of cases didn’t result in a conviction because it is problematic to prove that the person who spread the image had the intent to harm the victim. The fact that public prosecution was no longer dependant on victim reports also improves the capability of public prosecution to handle this phenomenon (Tweede Kamer der Staten-Generaal, 2019). It is possible that victims are unaware of the fact that they have been victimized because they are not yet aware of the existence of that pornographic material. For this criminal offense, the public prosecution will need to determine whether or not the manufactured content was made without the consent of the victim, but it is no longer necessary for public prosecution to have a report of that victim to continue with the case. This helps because victims can have several reasons not to file a report against the perpetrator. Reasons not to file reports are various, but examples of them are familiarity with the perpetrator, not knowing where, to who, and what they can report, and being afraid of repercussions (Jones et al., 2009; Spencer et al., 2017). Removing the necessity of a victim report for this criminal offense lightens the burden on the shoulders of the victims. They are no longer the ones who are the gatekeeper in the prosecution process.

This article can in my opinion be fruitfully used for the prosecution of cases in which deepfake technology is used for the creation of images of sexual nature against the will of the person that is portrayed in those images. This is also argued in the appreciation of the article (Berndsen, 2020). It is however only applicable in cases in which the created content is of sexual nature. In this case, there is jurisprudence in which someone secretly recorded a minor under the shower. The offender used those images to replace the face of different women with her face, creating different pornographic material. The offender used the software Photoshop to create the material and was found guilty of article 139h of the Dutch Penal Code among other offenses (ECLI:NL:RBDHA:2021:1885). The way in which Photoshop is used by the offender is legally comparable to the application of deepfake technology discussed in this paragraph, making it clear that article 139h of the Dutch Penal Code is applicable when deepfake technology is used to create pornographic material of a person.

The most difficult aspect of article 139h is determining what publication means for this article. When this was asked in parliament (Tweede Kamer der Staten-Generaal, 2019), the minister answered that by just showing the picture to others, or in some cases to one person, could be considered publication (Berndsen, 2020). This is defendable from the harm that can be inflicted on the victim, but this isn't clearly described in the article. To determine what the lower limits of this article are, we will have to wait for further jurisprudence (Berndsen, 2020). According to article 12 paragraph 1 sub 6 of the copyright law, "to publish" includes the disclosure by means of transmission of a work by cable or other means. This is legally comparable to cases in which "revenge porn" is being shared with others because that content is similarly transmitted by cable or other means.

Article 139h of the Dutch Penal Code is a relatively new law that regulates a specific kind of behaviour. According to Lessig's theory, the law can regulate both in a direct and indirect manner. This article is an example that regulates behaviour in cyberspace directly since it prohibits some applications of deepfake technology. Another interesting aspect, from the perspective of Lessig's theory, in this article is that it is very specifically tailored to a certain kind of behaviour. It is entirely focussed on the creation of pornographic material without the subject's consent. This is in line with what Lessig described in his theory that law directly regulates behaviour in cyberspace should be specifically tailored. Tailoring it specifically to pornographic material makes certain that it regulates only those specific cases. It is, however, not exclusive to cyberspace. It is possible to spread the media via other means than cyberspace.

It is also interesting to see that this article came into effect after our norms had changed. There was a need for a specific criminal provision regarding revenge porn and the norm of the general public was that existing laws didn't cover that phenomenon sufficiently. This norm then affects law, since it moved the legislative body to create new laws. This is an example where the modality of norms, from Lessig's theory, affects the modality of law.

Defamation

In general, when people have the feeling that they are being damaged in their name and honour by a deepfake, they have the option to sue for defamation. Defamation is an umbrella term for libel and slander. Libel has been criminalized in article 261 of the Dutch Penal Code and this states that whoever purposely damages someone's honour or name with the goal of making that public is guilty of libel and can be punished with a maximum of six months detention or a fine of the third category. If this is done by writing, then the sentence will be increased to a maximum of one year detention or a fine of the third category. Slander has been criminalized in article 262 of the Dutch Penal code and this states that if someone commits the crime of libel whilst knowing, or should have known, that the information is untrue. The main difference between the two is that for slander the information or media used is untruthful. With libel, it isn't necessary to prove that the information or media is untruthful. For both criminal offenses, it is important that the perpetrator knows, or should have known, that he is inflicting damage on the other.

These criminal offenses can be committed via deepfakes in the way that deepfakes can be used to purposely damage someone's name and honour. If deepfakes are being used in obvious nefarious ways, then these criminal offenses might be an option. However, it becomes difficult if the perpetrator challenges the fact that there is damage and that it is done on purpose. It is difficult to determine if someone published damaging content for someone on purpose or if he should have known that it is damaging to that person. There is a large grey area in which a judge will have to decide between the interests of both parties and this will be very dependent on the context.

It is however an option that victims of deepfakes have when they feel that the deepfake is damaging their name or honour. It is however important for that person to prove in what way the deepfake is damaging him or her. In some cases, it would be quite simple to prove that, especially in the case where deepfakes are being used to portray someone committing a criminal offense. In that case, it is quite clear that the name and honour of that person are damaged. But this becomes more difficult when the person who is portrayed in the deepfake is not directly, but indirectly damaged. This could be the case in examples where someone created a deepfake video of a politician participating in a certain demonstration while that politician in reality is not supporting the message of that demonstration. In that case, it is more difficult to determine if the name and honour of that person are damaged. Even if that politician loses the support and receives fewer votes in the election, it is almost impossible to determine if this is a result of the deepfake video. But it is clear that "fake news" in itself, and thus also manipulated media with deepfake technology portrayed as the truth, could have severe consequences on election results (Allcott & Gentzkow, 2017).

In the Netherlands, we have seen an example of this in 2017 when Geert Wilders posted a photoshopped picture of politician Alexander Pechtold. In this picture, the face of Alexander Pechtold was photoshopped in a picture of a pro-Islamic protest of Hamas saying that Pechtold was part of that protest (NOS, 2017). Wilders tweeted that picture in February which is a month prior to the lower house elections in the Netherlands. Pechtold decided not to file a complaint against Wilders with the public prosecutor. It would have been a perfect case in which the judge will have to rule on the boundaries of freedom of speech during election time and slander. This also shows that regular 'fakes' can have the same effect as 'deepfakes' the main difference is that with deepfakes, artificial intelligence is used to alter the imagery and photoshop still needs manual labour. But the results are similar and that's why case law on photoshop is interesting for this thesis.

This example is particularly interesting because in 2014 there was a case in which the judge had to rule on whether a cartoon of a lawyer was considered to be slander or satire. In that case, the judge ruled that it was slander because in the cartoon the lawyer was untruthfully portrayed as "shifty". The judge took into account that a cartoon needs to be interpreted in a way that the regular reader will interpret it. This means that the artist cannot expect the reader to delve into any deeper meanings to it. With the lack of background information, the portrayed cartoon can be interpreted as the truth, and therefore harmful to the lawyer (ECLI:NL:RBLIM:2014:9244). Comparing a cartoon to the photoshopped picture of Pechtold shows a few equalities. It is arguable that, as with cartoons, Twitter is a medium on which media is quickly consumed. Can it be expected of a reader of that tweet to delve into the deeper meaning of that tweet and to distinguish it from real and fake? If not, the judge could have ruled similarly in the case of Pechtold if he filed a complaint.

In summary, defamation could be used for certain applications of deepfakes, but it is difficult because the victim will have to prove that he or she has suffered damages. In some cases that is quite easy, but there is a large grey area in which a judge will have to decide based on the context of that case. Until now, there have not been any court cases yet in which someone was prosecuted for defamation by deepfakes. There have been other cases before in which there was a ruling on the freedom of speech and defamation, like the one stated before, and those cases are showing possible equivalencies with the use of deepfakes for defamation (ECLI:NL:RBDHA:2019:11523).

This article is, from the perspective of the theory of Lessig, also an example of the law that directly regulates behaviour in cyberspace. One could use deepfake technology to inflict someone’s reputation by spreading compromising media on the internet, thus via cyberspace. This article criminalizes that specific behaviour, making it a direct regulation.

Portrait rights

Deepfakes can also be evaluated from the field of portrait rights. In the case of deepfakes, faces and or voices are used of others to create images or other content. This creation can be classified as creating portraits of someone and the person which is portrayed does have certain rights which are enclosed in the copyright law in the Netherlands.

First is it important to distinguish if deepfakes can be considered to be portraits, because if they are not considered to be portraits, then the copyright law wouldn’t be applicable. At this time there is no jurisprudence about deepfakes yet, but there is jurisprudence that explains how a drawing of someone can be seen as a portrait if it contains characteristic features (ECLI:NL:RBAMS:2005:AS4748). The court ruled that even a drawing can be a portrait of someone and thus fall under the jurisdiction of copyright law. I would therefore claim that this is also the case for deepfakes. With deepfakes, you use an image or video of someone and the characteristic features of that person remain present. Therefore, I argue that deepfakes are susceptible to copyright law in the Netherlands.

The rights of the person who is portrayed are listed in article 21 of the copyright law. This fundamental article states that if a portrait is created without the order of the portrayed person, then publication is not permitted if that person resists that publication. In other words, the person who is on the portrait has to give consent to the publication of the portrait. If there is no consent, then the publisher violates article 21 of the copyright law and is punishable by article 35 of the copyright law which is an offense and its punishment is a fine of the fourth category.

An important notion here is that it is only an offense when it is being published. As discussed earlier, the definition of publication of copyright law is legally comparable, but it can be broadly interpreted. A clear example of this is the court case of Playboy Magazine versus GeenStijl. In that case, GeenStijl published a link to a website on which pictures that were the property of Playboy were unlawfully publicised. GeenStijl didn't host the website but did actively share the link on their website with the goal of making the content more accessible. In this case, GeenStijl was found guilty of violating the copyright law and they had to remove the link from their website (ECLI:NL:RBAMS:2012:BX7043). This shows that not only the publication of the content protected by copyright law is seen as an infringement, but posting a hyperlink to the protected content is also seen as such.

But, just making a portrait of someone without their consent isn't considered to be a violation of article 21, because that specifically states that it only focuses on portraits that are being published. Translating this to deepfakes concludes that making deepfakes of someone cannot be considered of violation of copyright law. But spreading that deepfake over digital platforms like Facebook or YouTube is considered to be a violation because then it is undeniable that the person is publishing the content.

In these kinds of cases, the judge needs to balance the different interests that are at stake. The person who is portrayed needs to state that they have a reasonable interest in the fact that his or her portrait cannot be used and the defendant needs to state why his or her interest is more significant. In practice, this might come to a judge deciding if the right of privacy of the person who is portrayed has more value than the right of freedom of speech of the defendant. If someone is portrayed as committing a criminal offense or doing something else that has a serious impact on his or her life, then the interest of the portrayed person will have more gravity than compared to a deepfake in which the portrayed person is just portrayed as doing nothing particular. This is a grey area in which a judge has to decide which interest is more important and this is always contextual. So solely the fact that someone published material without the consent of the subject does not immediately result in a conviction as seen in previous court cases (ECLI:NL:RBAMS:2019:7357).

At this moment there is no jurisprudence in which deepfakes are considered to be a violation of article 21 of the copyright law, but it does offer opportunities for the persons that are portrayed in deepfake videos or images. In another master thesis, Casper Heijnen evaluated the use of deepfakes with liability law in the Netherlands and concluded that it could be possible to receive compensation on basis of portrait rights when someone's image is being used to create deepfakes without their consent (Heijnen, 2021). It will be interesting to read what a judge will rule if such a case ever comes in front of a judge. That ruling might offer more criteria to better determine whether or not deepfakes are subjectable to copyright law in the Netherlands.

Portrait rights is a form of law that also regulates behaviour in a direct manner according to Lessig's theory on regulation in cyberspace. These rights are infringed when someone creates media with deepfake technology and spreads it via cyberspace without the consent of the subject. This regulation prohibits that specific behaviour by threatening the possible perpetrator with consequences, making it a direct regulation.

GDPR

The Netherlands is part of the European Union and this means that the General Data Protection Regulation (GDPR) applies. The GDPR has been in effect since 2018 and sets a benchmark for the processing of personal data. This GDPR states in article 5 that someone's personal data can only be processed if it is legitimate, fair, and transparent for that person. Processing is an umbrella concept for multiple different activities like storing, organizing, structuring, editing, and using. All of these actions can only take place if it is done legitimately, fairly, and transparently for that person. For it to be legitimate, fair, and transparent, the data subject must be aware of the fact that his or her personal data is used, that it is used fairly, and that the data subject has given consent for the processing of personal data.

The first important aspect here is to look at the term personal data. According to article 4 of the GDPR, personal data consists of any information relating to an identified or identifiable natural person. This entails all aspects that can be used to identify persons based on that data, i.e. distinctive tattoos. In the case of deepfakes, this personal data consists of images and video material with the whole purpose of impersonating that person. It is clear that it is possible to identify individuals with this data, making it personal data and therefore protected by the GDPR.

The GDPR states in article 3 that it applies to the processing of personal data of all data subjects who are in the European Union. It also applies if this processing is done by a controller or processor that is not established in the EU, as long as that processing is related to the monitoring of their behaviour in the EU. This means that people in the Netherlands are protected by the GDPR in the case of deepfakes even if the processor is based outside of the EU. This is the case because for deepfakes to be successful, the technology analyses the personal characteristics of that person to mimic behaviour in a video to be as lifelike as possible. There is also jurisprudence on this when a Dutch court ruled in 2020 that one cannot process personal data, including videos and audio, of someone without their consent (ECLI:NL:RBMNE:2020:24).

Importantly, Article 9 of the GDPR focuses on special categories of personal data. Article 9 states that processing personal data revealing biometric data is prohibited. Biometric data is personal data of specific technical processing of physical, physiological, or behaviouralist characteristics of a person. An image or video does display physical and behavioural characteristics of a person and therefore can be considered to be biometric personal data (Autoriteit Persoonsgegevens, 2021). There are exceptions in which this biometric personal data can be used, like for example access security, but only with the consent of that subject. In the case where someone is making a deepfake of someone else, using biometric personal data, it will not be with the consent of the data subject and therefore unlawful.

When your personal data is being unlawfully processed, you could then report this, according to article 77 of the GDPR, to a supervisory authority which in the Netherlands is the Autoriteit Persoonsgegevens (AP). When organizations are exploiting your personal data without your consent, then the AP will conduct research and that organization might receive a fine of a maximum of 20 million euros or 20% of their global revenue. Besides that, the data subject could also, according to article 82 of the GDPR, file for compensation for the damages that the data subject has suffered.

One important aspect of the GDPR is that it is focussed on processing personal data for commercial purposes. When handling personal data, the GDPR is applicable, with the exception if it is used in the household. This means that processing the personal data of your family at home is not covered by the GDPR.

One might look into the organisations that make the technology of deepfakes available for some kind of civil liability to compensate for damages inflicted by the software. But I think that this will not hold in court since there is no case yet in which organisations that are making the technology are held accountable for the content that is being created with that technology. It is the same with the manufacturer of crowbars. Crowbars can be used to break into homes, but that doesn't make the manufacturer of crowbars liable for every burglary in which a crowbar is used to gain entry to the house. This is of course a very simplified example to which there are a lot of nuances. This has been shown recently with the Israeli company NSO Group's product called ForcedEntry (CyberPeace Institute, 2021). This product is a very advanced version of spyware that has been sold to different states worldwide. There are calls to start regulating the manufacturing of spyware technology and this could also be a possible future for deepfakes if the problematic applications become more excessive and start to overshadow the positive applications. An explanation for this call for regulation might originate from the fact that spyware has no benign purposes. It is purely focussed on violating someone's privacy by monitoring their activities. Thus, when the problematic applications start to suppress the positive applications, this call for regulation might affect deepfake technology as well.

A solution might be to create policies that deepfake technology is not available for individuals. Making it only available for companies or similar entities ensures that the data used is also under the protection of the GDPR. And when looking at the problematic applications of the technology, most of them are when individuals are targeting different individuals. Creating a barrier, or at least making it more difficult, for individuals to use this technology might have a limiting effect on the number of problematic applications of deepfake technology. The main issue with this solution is the question of enforcement. Is it possible to fruitfully ban consumers and individuals from using deepfake technology? Personally, I don't think that this will be fruitful since the technology is already available to the public. Plus, I believe that it is better to regulate the correct use of technology instead of banning it.

The GDPR has been incorporated into Dutch domestic law and can regulate behaviour both in direct and indirect ways as discussed in Lessig's theory on regulation in cyberspace. Everyone that processes personal data, has to comply with the GDPR and is therefore regulated by its provisions. However, the GDPR also affects architecture in cyberspace. Since it is not allowed to process personal data without consent, it was necessary for the architecture in cyberspace to adapt to comply with the GDPR. This regulation is different from the other regulations because this also directly targets digital services in cyberspace. Not only the people who use certain services are bound to the GDPR, most of the time companies are making a profit from analysing and selling personal data as well. With the GDPR, these companies had to change their architecture to either make sure that they have the consent of their subjects or not to process personal data at all.

Analysis

After evaluating all these laws, it is interesting to examine if there is a difference between them concerning publication and creation, the role of intent, and if there are any common denominators. This is important because it will reveal underlying similarities and possible gaps in the regulation of deepfake technology.

Article 139h of the Dutch Penal Code criminalizes both illegal publishing and creation of images of sexual nature. This means without the consent of the subject on that image. Publishing is even an aggravating circumstance that increases the possible punishment from one to two years of imprisonment. With defamation, the main focus is on publishing since it criminalizes the spreading of discrediting information. Here it is not important who created the information and the information might even be true in nature, but the criminalization of this behaviour comes forth from the fact that it is disclosed in public with the goal of harming someone. The same can be stated from the subject of portrait rights. It is allowed to create pictures of someone for your own use, but publishing them requires the authorization of the subject. This means that the offense in portrait rights lies in the fact that it is published. The GDPR has no aspect of publishing or creation. It is mainly focused on how personal data is used and processed.

Publication and creation

It is interesting to see that there is a difference between the action of publication and creation when it comes to the problematic applications of deepfake technology included in this thesis. Creation is only regulated in article 139h of the Dutch Penal Code and the GDPR. The others are focussed on publication instead of creation. The reason for this might be that there is something called artistic freedom. Creating images of someone in a compromising setting does not immediately harm that individual. It might inflict harm when it is published, but until that is the case, purely the creation might not result in any harm for the subject. Article 139h of the Dutch Penal Code is different in this case because it is focussed on sexually oriented imagery. With this kind of content, which can inflict serious harm to victims, the government wanted to do right to that harm and therefore it is treated differently than defamation.

With portrait rights, the whole goal is to protect the rights of a subject concerning the publication of its image. Therefore, it is purely focussed on publication instead of creation. It is not illegal to use deepfake technology to alter images of someone, as long as it is not published without the consent of the subject.

The GDPR does not have a provision for publication. It is purely focussed on processing personal data. As argued earlier in this thesis, using deepfake technology to alter the imagery of someone is considered to be processing personal data. Therefore, the GDPR does cover the creation of imagery using deepfake technology, but only as long as it is used to create or alter imagery of someone. This is the case because it is only applicable to personal data. For instance, using videos of someone to create new videos on which that person is doing something else, is protected by the GDPR, and therefore it is applicable to the problematic applications as discussed in this thesis. Interesting about the GDPR is that harm is considered to be present by just using the subject's personal data. When individuals are using personal data for their own use, i.e. a picture, it isn't considered to be illegal, with the exemption of sexually oriented images protected under article 139h of the Dutch Penal Code.

Intent

In article 139h paragraph one sub a of the Dutch Penal Code, intent is explicitly named as one of the components of the criminal provision and includes all degrees of intent including conditional intent. Conditional intent is still intent in the Dutch legal system, but it is a broader definition than solely intent. With conditional intent, the perpetrator could and/or should have known of the consequences (Buruma, 1999). In other words, the perpetrator acted recklessly, and even if the perpetrator didn't intend that specific outcome, by acting recklessly he or she accepted the reasonable chance of this specific outcome. Here it is clear that intent is necessary for a criminal conviction. Even though intent is only explicitly mentioned in paragraph one sub a, it is implicitly present in the other paragraphs as well (Berndsen, 2020). Making something public is an action in which intent is included since it also includes conditional intent.

Intent is not present in portrait rights as an element of the law. It is however disclosed in the concept of publishing and therefore it is implicitly present in portrait rights (Darras, 2011). When the perpetrator published an image of someone, the act is considered to be intentional. The perpetrator might not have the intention to harm the subject, but there is an intent for publishing the material.

Common denominators

Besides publication, creation, and intent, some other concepts are present in the laws that are discussed in this thesis. One important component is the concept of harm. In article 139h of the Dutch Penal Code and with defamation, harm is considered to be present since it inflicts damage to the victim. The images created with deepfake technology are used to inflict harm like reputational damage. Portrait rights and the GDPR consider there to be harm by solely the fact that someone's image is being used. This shows that harm is different in criminal law than in portrait rights or the GDPR which is due to the fundamental difference between criminal and civil law.

The final important component of all these laws is the concept of consent. The scope of this thesis is restricted to illegal forms of deepfake applications which implies a lack of consent. This concept of consent is explicitly present in all these articles, making it clear that it is allowed to use deepfake technology for the problematic applications discussed in this thesis, as long as the subject has given consent. It might be interesting to see the extent of consent and in what form consent needs to be present, but that isn't included in this thesis.

Analysis summary

Analysing the different concepts of the laws regarding deepfake technology, allows us to place it in an overview as shown in table 1. This overview shows that there is a difference between these laws regulating how deepfake technology is being used with regard to problematic applications. Article 139h of the Dutch Penal Code has the most applications because it is specifically focussed on sexually oriented images. Given the specific nature and serious consequences of that kind of media, it is apparent that the government decided to criminalize and include more concepts in comparison with defamation, portrait rights, or the GDPR.

	Publication	Creation	Intent	Common denominators
Article 139h Dutch Penal Code	X	X	X	Harm, consent
Defamation	X	-	X	Harm, consent

Portrait Rights	X	-	X	consent
GDPR	-	X	-	consent

Table 3: Analysis overview of different laws regulating deepfakes

Conclusion

This chapter discussed the different laws that might apply to problematic applications of deepfake technology. It shows a different array of laws, in which the main focus was on the Dutch Penal Code, portrait rights, and the GDPR. I have argued that these laws can be applicable and it is clear that the current legal framework offers several options to tackle these problematic applications. It is difficult to determine how it would play out in court since there is no jurisprudence yet in which deepfake technology played a major role. The fact that there is no jurisprudence yet can be a confirmation of the fact that this technology is relatively new. It can also be a sign that prosecution is too difficult for cases to make into court. This will be briefly discussed in the discussion chapter of this thesis and might be an interesting subject for future research.

The laws that are discussed in this chapter are not an exhaustive summary of the laws that apply to deepfake technology. This thesis focussed on these because these laws are, in my opinion, the most applicable at this moment. I believe that they are the most applicable since these laws are mostly used in jurisprudence on similar cases. There is no specific law for deepfakes, so general laws must be consulted on applicability. These laws have been used in similar cases in which media has been altered. The only difference with them is that the alteration or creation is done with deepfake technology. This makes these laws the most applicable to deepfakes at this moment. In the future, there might be changes in the legal landscape, making it necessary to keep evaluating how the application of deepfakes fits in that legal framework. It is also possible that there will be jurisprudence in the future which will define whether or not these laws apply to those applications of deepfake technology, also making it necessary to evaluate the legal framework.

Evaluating these laws from the perspective of Lessig's theory has shown that there are certain laws that have a direct effect on behaviour in cyberspace. The GDPR even has an indirect effect since it affects architecture in cyberspace that affects the use of personal data. It is interesting to see that there the laws that have been evaluated are general in their use and not cyberspace specific. I expected that law would be general and not cyberspace specific, and this evaluation has confirmed that expectation. I also expected that the modality of law could have a large effect on behaviour and after evaluating these different laws I must conclude that the legal framework offers several options for the problematic applications of deepfake technology.

Since there seems to be no immediate gap in the legal framework for these problematic applications of deepfakes, it is important to look at different regulations that could regulate the use of deepfake technology. The next chapter will focus on online content regulation and evaluate who has which responsibilities in the landscape of digital platforms.

Deepfake regulations on digital platforms

Besides looking at the individual who makes or spreads the deepfake material, one might also look at the platforms that are being used to share that content. To narrow it down, this paper focusses on the legal regulations to which social media platforms have to adhere. For future research, it will be interesting to have a broader examination of regulations regarding social media platforms so it can include a broader perspective than formal regulations.

A lot of deepfake material is being spread via the internet and in particular via social media. This fuels the discussion on to what extent social media platforms are responsible for the content that is posted by their users. Multiple studies have been conducted in this area and all of them conclude that social media platforms can play a role in regulating misinformation (Carlson, 2018; Chesney & Citron, 2019; Chow, 2019; O'Donnell, 2021; Reisach, 2021). This can vary from having responsibility for alerting its users for misinformation to removing the content. At this moment there are even studies and initiatives to create liability with the social media platforms if they fail to remove or prevent the dissemination of misinformation (O'Donnell, 2021).

The question of liability for intermediaries has been evolving in the last few decades and the foundation was provided in 2000 with the EU Directive on Electronic Commerce. Article 14 states that intermediaries weren't liable for the hosted information as long as they aren't aware that the content is illegal, or that when the intermediary becomes aware of the illegal content, promptly deletes or denies access to that content (Jørgensen & Zuleta, 2020).

However, article 15 of the directive states that there is no general monitoring obligation for intermediaries. This meant that, in practice, intermediaries would only become aware of illegal content when that content is being addressed by their users. And to prevent liability, intermediaries would address the reported content by its users, but they shall not actively monitor its users. The problematic aspect here is that it is not clear how quickly the intermediary is expected to act on illegal content (Riis & Schwemer, 2019).

While this directive was the main regulation for intermediaries for seventeen years, different regulations started to emerge in 2017. The European Commission issued a communication called "Tackling illegal content online" in 2017. This was aimed at creating "enhanced responsibility of online platforms" (European Commission, 2017). Following this, the European Commission issued a recommendation on "measures to effectively tackle illegal content online". This recommendation aimed to encourage hosting service providers to take appropriate, proportionate, and specific proactive measures concerning illegal content (European Commission, 2018a). The interesting here is that the recommendation is focussed on proactive measures. This implies that the hosting service providers will have to actively monitor their users if they wish to fulfil the recommendation of the European Commission. This recommendation for proactive measures was also present in the Terrorist Content Regulation in article 6 (European Commission, 2018b).

These regulations make it difficult for hosting service providers like social media platforms to comply. The EU Directive on Electronic Commerce states that social media platforms should not actively monitor their users, but that, when aware of illegal content, have to take measures to remove the illegal content to benefit from the "safe harbours" provisions and thus prevent the liability of that content. But in contrast, they are expected to undertake proactive measures according to the Terrorist Content Regulation and recommendation of 2018. And third, it also needs to take human rights into consideration concerning the freedom of speech and access to information (Jørgensen & Zuleta, 2020).

Additionally, the European Union has recently adopted a new Digital Service Act (DSA) which will be implemented in 2022. In the DSA there are obligations to digital platforms to protect the users. The first obligation is stated in article 26 and entails that digital platforms will have to assess any risk of the dissemination of illegal content from the functioning of their services (Savin, 2021). After making these risk assessments, the platforms are obligated according to article 27 to take "reasonable, proportionate and effective" measures to mitigate those risks. These articles are extending the responsibility of the digital platforms since this entails that the digital platforms will have to take more proactive measures regarding illegal or unlawful content. The DSA will have to be implemented in domestic law, but with the adoption of the DSA, Europa has set an important step into making digital platforms responsible for proactively monitoring their content.

This DSA is interesting because it is a legal regulation that will force digital platforms to change their architecture. Since it is impossible to manually check all the content on their platforms, they will have to implement the provisions of the DSA in their code or algorithms. This is a perfect example in which the modality of law affects the modality of architecture from Lessig's theory on regulation. Lessig stated that law can have a large effect on behaviour by affecting architecture, and the development of this regulation shows that the European Union invests in indirect regulations.

A recent study shows that the majority of people think that social media platforms are responsible for preventing the dissemination of misinformation. One of the findings of this study is that there is a large social demand for accountable social media platforms that are actively preventing the spread of misinformation (Lima et al., 2022). This means that the public has a positive attitude towards giving social media platforms more responsibilities in the public debate. If social media platforms are responsible for censoring content in the public debate, this might result in an infringement of the freedom of speech (Chesney & Citron, 2019).

In the Netherlands, some regulations apply the problematic applications of deepfake technology. It is stated in the current guidelines for electronic trade that hosting service providers aren't liable for the information they host if they have no knowledge or reasonably should know that the information that they are hosting is illegal or unlawful (Ministerie van Justitie en Veiligheid, 2020). The interesting aspect here is that hosting providers are by default not liable. Only if they have a notion that they are hosting illegal or unlawful information makes them liable. This means that it is important for the hosting providers to have a system that allows them to act fast when someone notifies them that they are hosting illegal or unlawful information (Schermer & Sloot, 2020). In the Netherlands, there was also a court ruling that stated that hosting providers don't have to actively monitor their services. However, it can be expected of them that they will conduct their activities with socially responsible care (ECLI:NL:RBAMS:2019:8415). The downside of this is that the term "socially responsible" isn't elaborated further and is open for interpretation. This leaves a grey area in which organisations have to operate and that could lead to unwanted consequences like censorship. And with the emergence of the DSA, this might become overruled and give organisations a responsibility to actively monitor the content of their services. This will become clear in the coming years when the DSA has been incorporated into Dutch domestic law.

All of these regulations are focussed on illegal content. The problematic applications of deepfake technology, as discussed in this thesis, are illegal content. They are illegal since they are examples that are infringing several laws, including the laws that are included in this thesis. So, when digital platforms like social media must undertake measures against illegal content, then it will include the problematic applications of deepfake technology.

This becomes increasingly interesting with the implementation of the DSA. Digital platforms will have the obligation to make a risk assessment on illegal content dissemination and will therefore have to make this assessment on deepfake content as well. But how will they be able to determine if the content is illegal? This is incredibly difficult, if not impossible, for digital platforms since they will have to determine if the content is made with the consent of the subject. This has the consequence that digital platforms will need to implement some form of detection measures that will help determine if the content has been altered or created by artificial intelligence. And even if the content has been altered by artificial intelligence, will the digital platforms be able to determine if that content is illegal? This puts digital platforms in a precarious position because they need to undertake measures to mitigate the risk of illegal content.

In the DSA it is stated that platforms have to make a risk assessment and then take appropriate measures to mitigate those risks. Looking at the feasibility discussed in the previous paragraph, it is clear that it will not be possible, or hardly possible, to mitigate all risks of illegal content created by deepfake technology without completely banning deepfake technology as a whole. It is also not desirable to ban all applications of deepfake technology since it will also ban the positive applications as discussed in this thesis. So, what are appropriate measures to mitigate those risks?

I would argue that appropriate measures are measures that do right to the possible harm of problematic applications but will allow the continuance of positive applications of deepfake technology. The most harm by imagery altered by deepfake technology is that people are not able to determine whether the imagery is altered or fake and therefore accept that image as reality. If people don’t believe it is real, then there will be less inflicted harm to the victim. So, a solution might be that digital platforms will implement software to detect if the imagery is altered or created with deepfake technology and then flag that content. This will create a situation in which the public will know that the image portrayed is not real and that will result in less harm. This is also in compliance with Lessig’s theory on regulation because he states that behaviour in cyberspace can effectively be regulated by law affecting the architecture. Implementing this solution as a mitigating measure that is obliged by the DSA makes it another example of the modality of law affecting the modality of architecture.

This option is not infallible since it is still possible that people will believe that the content is real or that different aspects of the image have been altered. This might be mitigated if the software didn’t only flag the content as altered, but could also identify what part of the image has been altered. There is software available that can detect if deepfake technology has been used to alter imagery and this could be an effective measure for digital platforms to comply with the DSA. It would still enable them to be a platform on which positive applications aren't banned, limiting the effect on human rights such as freedom of speech and access to information, and reducing the harm inflicted by problematic applications of deepfake technology.

The EU has also drafted an Artificial Intelligence Regulation in which deepfakes are named explicitly. It is still a draft and might take a while for it to be implemented into domestic law, but the proposal states a new transparency rule for the use of artificial intelligence. Users of artificial intelligence that create or alter content are obliged to mention that it has been artificially modified (European Commission, 2021). Article 52 sub 3 mentions that users of an AI system that generates or manipulates media that resembles existing persons, objects, places, or other entities or events and would falsely appear to be authentic, shall disclose that it has been artificially manipulated. However, it will not apply to deepfakes if it is necessary for the exercise of the right to freedom of expression. This regulation is interesting because it tries to directly regulate behaviour in cyberspace by giving users of deepfake technology the obligation to mention that the content has been artificially modified. This is a direct effect of the modality of law on behaviour in cyberspace.

This regulation could be a powerful combination with the DSA since it combines direct and indirect effects of law on behaviour according to Lessig's theory. Since the DSA could lead to digital platforms modifying their architecture, the indirect effect of law, makes it complementary to the Artificial Intelligence Regulation, the direct effect of the law on behaviour.

This proposal received several comments and one specifically on the exemption of deepfakes. The Commission Meijers proposes that the European Commission should rethink this exemption because it conflicts with the entire provision. Since deepfakes are a communicative act, most deepfakes will fall under the exemption of the right to freedom of expression (Meijers Committee, 2022). In other words, almost all deepfakes will fall under this exemption and the regulation will therefore have little to no effect on problematic applications of deepfakes. It will be interesting to see if this part of the AI Act will be revised and how this provision will be translated into domestic law.

Interestingly enough, social media platforms have created policies to regulate the spread of deepfakes on their platforms. Meta (parent company of Facebook and Instagram) states that it will remove manipulated media, e.g. deepfakes, where it could mislead others (Europol, 2022; Meta, 2022). TikTok has implemented a ban on altered or synthetic media that could mislead users (Europol, 2022; TikTok, 2020). Reddit bans content that impersonates individuals or entities presented to mislead others (Europol, 2022; Reddit, 2020) and YouTube has an existing ban for manipulated media in their community guidelines (Europol, 2022; Google, 2022). But these are just policies and they are only enforced when the social media platforms are notified that content violates their policies.

Additionally, many of the policies do not block the use of deepfake technology for satire and this means that intent becomes an important aspect again. This aspect of intent is extremely difficult to determine solely on the content and would therefore maintain the uncertainty of the level of effectiveness of these policies (Europol, 2022).

Besides these policies, some companies like Meta, Google, and Microsoft, have been developing software to detect deepfake technology which could be used to enforce their policies in the future (Europol, 2022). The future will tell us if those techniques are effective, or whether they will remain incapable of defining intent and therefore incapable of enforcing the policies as they are at this moment.

Regulation theory of Lawrence Lessig

As discussed in the methodology chapter, this thesis will look at the different regulations from the perspective of Lawrence Lessig's theory of cyber regulation. In that theory, Lawrence Lessig states that behaviour is regulated by four different modalities: laws, social norms, markets, and architecture (Lessig, 1999). These different modalities do not operate independently, but they are interacting with each other, i.e. social norms in society can determine what is available on the market and that will affect what laws are possible. He states that behaviour would mainly be regulated by law and architecture with both direct and indirect effects. An example of direct effects is how the law regulates behaviour because people fear repercussions from law enforcement. An example of indirect effect is how the law regulates the architecture of certain applications, like the prohibition of discrimination in algorithms, which in turn affects the individual using that application. Both effects can be used to steer behaviour in a certain direction and therefore regulations (Lessig, 1999).

This thesis has looked at several laws that directly and indirectly influence behaviour in cyberspace. It is in line with the expectation of Lessig that most regulations will be focussed on law and architecture as these were the regulations included in this thesis. However, the other modalities can affect behaviour in cyberspace as well and those will now be further discussed.

Social norms

Social norms can regulate behaviour because people can experience negative consequences from their social environment. If your peers disapprove of the problematic applications of deepfake technology, you are then less likely to engage in that behaviour. Now, deepfakes are often seen as a gimmick, but the rise of problematic applications might change the attitude of the general public towards deepfake technology.

Often laws emerge from norms in society, but laws can also have an effect on the moral attitude of the general public (Aksoy et al., 2020; Kotsadam & Jakobsson, 2011). This means that when the government implements laws concerning the use of deepfakes, it has an effect on the attitude towards deepfakes. Article 139h of the Dutch Penal Code, which criminalises "revenge porn", could therefore change the attitude towards the creation of media of sexual nature by deepfake technology without the subject’s consent. If this application of deepfake technology increases in frequency, it might affect the attitude towards deepfake technology in general.

Markets

Markets can regulate behaviour by affecting the availability of software that both creates and detects deepfakes. With the new EU regulations, digital platforms need to make a proper risk assessment concerning the spread of illegal content. And digital platforms will have to take measures to mitigate the risks that are identified in that risk assessment. This means that, as discussed in this thesis, digital platforms will have to take measures to mitigate the spread of illegal deepfakes. Detection software can play an important role in mitigating measures for digital platforms. The balance between creation and detection software might be changed if detection software becomes more interesting as a means to mitigate those risks. Governments can also impose changes in regulation that make lower the costs or increase the benefits of detection software, facilitating increased self-regulation (Lessig, 1999). This could create incentives for companies to engage with detection software.

Architecture

Architecture can have a large regulatory effect on behaviour in cyberspace (Lessig, 1999). In cyberspace, it consists mostly of codes and algorithms that are designed with a specific purpose. Changing the architecture of applications or platforms can have a direct effect on behaviour because it might limit the capabilities of that platform. It is possible to create code that denies the publication of deepfakes, therefore regulating behaviour of the person that tries to publish it.

With the implementation of regulations like the DSA, digital platforms will have to implement measures to mitigate the spread of illegal content. Detection software can be implemented on digital platforms, making it a part of the architecture of that digital platform. This software is not yet implemented on all our devices or services, but it might be an effective measure against the problematic applications of deepfake technology because it will reduce the plausibility of that content.

Conclusion

Research aims and questions

The aim and research question of this thesis was to make a legal analysis of the applicability of regulations on deepfake technology in the Netherlands and to determine if there are any gaps that need to be addressed. To answer this, this thesis has described what deepfakes are, some problematic and positive applications of deepfake technology, and how deepfakes can be identified. After that, this thesis reviewed several laws and their applicability to deepfake technology, specific to the problematic applications. To make it complete, this thesis has also reviewed present regulations on digital platforms, like social media, with regard to deepfake technology. These different regulations were viewed from Lessig's theory on cyber regulations. By looking at all these different aspects, it is possible to review the regulations that are applicable to deepfake technology in the Netherlands.

Findings

Deepfakes are media that are altered with the use of artificial intelligence. Artificial intelligence is being used to identify facial movements in original media and it can use that data to create new or altered media. It is done via deep learning called Deep Neural Networks (DNNs) and Generative Adversarial Networks (GANs) which are a network of a large compilation of artificial neurons that do simple computations. When these neurons are working together, they can solve complex issues like re-enacting facial expressions.

Some applications of this software that are arising are troublesome. This thesis reviewed the problematic applications of deepfake technology in identity theft, pornography, sabotaging others and elections, and damage to the trust of people or organisations. More positive applications are the use of deepfake technology in mourning therapy in which people are able to converse with deceased loved ones, coping with disabilities like the loss of vocal cords, in education, it could offer new ways to engage students with the subject matter, and it can also be used to test hypothetical situations like having a different hairstyle.

It is possible to identify deepfakes using the same technology to create deepfakes. With DNNs and GANs, it can be determined if the media is altered by deepfake technology. However, it is still evolving and not widely available. The software to create deepfakes is more advanced and readily available than software to detect deepfakes which will remain an issue for the future.

There are regulations that try to mitigate the problematic applications of deepfake technology in the Netherlands. Article 139h of the Dutch Penal Code criminalizes the creation or publication of media of sexual nature without their consent. The analysis suggests that this article is applicable to media that is created or altered by deepfake technology. Interesting about this article is that it removes the intent to harm the victim, just the act of creating or publication is enough for a criminal offense. It is, however, difficult to determine what the lower limits are of the term to publish in this article and we will have to wait for jurisprudence for this to be set.

Defamation is a criminal offense that might be applicable to problematic applications of deepfake technology. Deepfake technology can be used to create media of someone that harms someone's name and honour and could therefore be used as an option by the victim to report the creator or publisher to the authorities. The problem with this option is that the victim will have to prove the damages that are inflicted because of the deepfake media. And the judge has to balance freedom of speech with the possible inflicted damages.

Deepfake technology uses images of someone to create or alter different media and is therefore subject to portrait rights. There is no jurisprudence yet in which portrait rights are used to receive compensation for illegally created deepfakes, but the findings suggest that article 21 of the portrait law does offer an opportunity to gain compensation for the victim.

The GDPR is a stranger in the midst since it does not target the publication of deepfakes. It is purely focussed on processing personal data. It is applicable to deepfakes since media that includes facial characteristics, like images, are considered to be biometric data and thus personal data. The GDPR regulates the processing of personal data, thus if deepfake technology is used to process personal data, it will have to comply with the GDPR, making it that it can only be created if the subject has given consent.

Interestingly, there is a difference between creation and publication. In most regulations, the focus is on publication instead of creation. This means that to successfully regulate deepfakes, one must look at what barriers are present in publishing deepfake media and therefore look at the regulations that are applicable to digital platforms on which media is being published, like social media.

Intent is a common denominator that is present in all regulations, except the GDPR. This means that it is important to establish that accidental violations of those regulations, do not immediately have to result in compensation. However, there is conditional intent which includes intent if the offender could and should have known that harm would be inflicted.

In the past, there haven't been a lot of regulations that were present on digital platforms and their content. Most of the regulations resulted in a lack of liability for the digital platforms, resulting in a lack of incentives for the digital platforms to implement any regulation at all. But several EU directives and recommendations have led to the appearance of the Digital Service Act (DSA) in 2022 which gives digital platforms the responsibility to actively monitor the content on their platforms. This will need to be translated into domestic law for it to become an effective regulation in the Netherlands. Until that is done, companies are operating in a grey area where they are obligated to remove illegal content, but are not obligated to actively monitor their content. Which makes it an ineffective regulation since the harm could already be done before the content has been deleted.

The research also suggests that we are moving towards a situation in which private companies, like digital platforms, are the gatekeeper of what content is allowed. This is interesting because this means that they will infringe on our human rights like freedom of speech and therefore have to balance the right of its users with the obligation to prevent illegal content on their services.

Interpretation

These findings answer the first part of the research question since there are several regulations applicable to deepfakes in the Netherlands. Two articles of criminal law have been reviewed and found to be applicable to problematic applications of deepfake technology, namely article 139h and defamation. Both of these articles are applicable to the problematic applications of deepfake technology as discussed in this paper. These are also the most important regulations since the Dutch Penal Code allows law enforcement to investigate and collect evidence. This is different from the civil law regulations that are included in this thesis. With portrait rights and the GDPR, one can claim compensation, but the victims must file a case themselves and bring it to a judge.

Besides these laws on the behaviour of individuals, there are also regulations on the digital platforms that are used to share deepfake media. It is interesting to see that there have been regulations for a while, but they were not obligated to take proactive action against illegal content on their services.

With the coming of the DSA in 2022, this might change. But we must ask ourselves the question of how these digital platforms will implement this obligation for proactive monitoring. If digital platforms don't take sufficient action, then they might be opening themselves to liability. This might lead to situations in which digital platforms will overcompensate and will possibly censor more content than necessary.

It is clear that it is necessary to have regulations that tackle the problematic applications of deepfakes. However, it would not do right to the positive applications of deepfakes to outright ban the technology. There should be a balance in regulation that does focus on the problematic applications of deepfakes but should leave enough room for the technology to be used for positive applications. That is why the positive applications of deepfakes are also an essential aspect to consider when reviewing regulations regarding deepfakes.

There are also some gaps that have been identified in this thesis. The first gap that has been identified is that there is no lower limit for the concept of publication incorporated in article 139h of the Dutch Penal Code. This means that it is not yet clear what the exact definition of publication is for this criminal offense. This will become clear when there is more jurisprudence since the judge will have to rule on different cases with different forms of publication.

The second gap is that it is unclear how the development of the DSA is going to continue when it will be translated into domestic law. It will obligate digital platforms to take measures to mitigate risks that become apparent when they have executed a risk assessment on illegal content. This is a gap because it is unclear how this will be implemented in the future. It could result in private companies becoming the gatekeeper of our freedom of speech because they will decide what can be shared or not on their digital platforms. In my opinion, that should always be decided by a democratic process and should not be solely decided by private companies. I believe this is the case because I believe that the public has to decide what is right and wrong and not private companies that are, mostly, driven by profits.

This interpretation has helped the field of research by reviewing the different laws that are applicable to a relatively new phenomenon called deepfakes and it also identified two gaps that can be addressed in future research.

Lessig's theory has been used to evaluate the regulations in this thesis. As Lessig explained in his theory, the most effective modalities to regulate behaviour in cyberspace are law and architecture. The regulations that are included in this thesis support Lessig's theory that the main focus is on these two regulations. The most current forms of regulations are from the modality of law and architecture. As shown in this thesis by the different regulations, the modality of law can affect the modality of architecture making it possible to not only directly affect behaviour but also in an indirect manner.

The most current regulations, like the DSA and the Terrorist Content Regulation, show that these two modalities are the main focus of governments to try and regulate behaviour in cyberspace. This supports the theory of Lessig that claimed that those two modalities are the best option to try and regulate behaviour in cyberspace. We will have to wait and see how effective these regulations are because some of them still need to be implemented and there is no jurisprudence yet, but the current developments show that there is hope and belief that it is possible to effectively regulate behaviour in cyberspace.

Most of my expectations have been confirmed in this thesis. I expected that the biggest effect on behaviour in cyberspace would come from the modality of law and architecture. This is supported by the fact that most recent forms of regulation are focussed on the modalities of law and architecture. Another expectation I had, was that the modality of law would have both direct and indirect effects on behaviour in cyberspace. The different regulations in this thesis confirm this expectation. My expectation that cyberspace would not be well regulated was not fulfilled. This thesis has shown that there are many regulations present in cyberspace. The main aspect however is that these regulations are general in nature. There are few to no regulations specifically for the problematic applications of deepfake technology. But the fact that there is an array of different regulations still applicable to those applications, shows that there is no lack of direct regulation on behaviour at this time.

In summary, this thesis researched an array of regulations that might be applicable to the problematic applications of deepfake technology. It consists of both regulations for the individual as well as regulations for digital platforms, like social media. The thesis also recommends further research on the implementation of the DSA since the implementation of the DSA will determine the balance of freedom of speech and censorship.

Recommendations for future research

Future research could focus on several aspects of this thesis. The first aspect future research could dive deeper into is the lower limit for the concept of publication that is present in article 139h of the Dutch Penal Code. It is possible to just wait for jurisprudence to see what the judge will rule as the lower limit, but that is circumstantial. Findings of future research on this aspect might help the magistrate to better determine the lower limit in its ruling. It might also help to determine the exact lower limit without having to wait for jurisprudence.

Subsequently, this thesis recommends monitoring the implementation of the DSA in Dutch domestic law. Since it will obligate companies to take measures to mitigate the risk of illegal content, this might result in censorship. Future research might help to determine how far the implementation of the DSA can go without infringing the right of freedom of speech.

Future research could also evaluate if there are differences in regulation between the Netherlands and the different countries. Since deepfakes are relatively new and every country encounters these problematic applications, it might be interesting to see if we can learn from different countries.

The last recommendation is to research the possibility of a content policy based on flagging content that has been altered or created with deepfake technology. Although this is a whole new research area, the findings of my thesis on the applicability of regulations of deepfakes and the possible gaps, suggest that flagging could be useful. Flagging might reduce the harm inflicted by the problematic applications because it reduces the chance that the public will consider the content to be authentic, while it will limit the effect on basic human rights such as freedom of speech. It could be researched in the future if this proposed policy will have the desired effect or not. This research will also help to identify the possible effect of the AI Act since that will entail a transparency obligation for users of artificial intelligence to disclose that the content has been altered with artificial intelligence. This disclosure is similar to the flagging proposal in the sense that it is clear to the public that the content has been altered so that it is clear that the content is not authentic.

Bibliography

Aksoy, C. G., Carpenter, C. S., De Haas, R., & Tran, K. D. (2020). Do laws shape attitudes? Evidence from same-sex relationship recognition policies in Europe. *European Economic Review*, *124*(11743). https://doi.org/10.1016/j.euroecorev.2020.103399

Alexander, E. C., Mader, D. R. D., & Mader, F. H. (2019). Using social media during the hiring process: A comparison between recruiters and job seekers. *Journal of Global Scholars of Marketing Science*, *29*(1), 78–87. https://doi.org/10.1080/21639159.2018.1552530

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Argos. (2022). *Deepfakes zijn niet meer weg te denken uit de informatieoorlog*. https://www.vpro.nl/argos/lees/onderwerpen/artikelen/2022/deepfakes.html

Autoriteit Persoonsgegevens. (2021). *Biometrie*. https://www.autoriteitpersoonsgegevens.nl/nl/onderwerpen/identificatie/biometrie

Bates, S. (2017). Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors. *Feminist Criminology*, *12*(1), 22–42. https://doi.org/10.1177/1557085116654565

Berndsen, M. (Michael). (2020). Een verbod op wraakporno. *Nederlands Tijdschrift Voor Strafrecht*, *1*(2), 70–76. https://doi.org/10.5553/nts/266665532020035002003

Buruma, Y. (1999). *Strafrecht is schuldstrafrecht*. https://repository.ubn.ru.nl/bitstream/handle/2066/128332/128332.pdf

Caldwell, M., Andrews, J. T. A., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. *Crime Science*, *9*(1), 1–13. https://doi.org/10.1186/s40163-020-00123-8

Carlson, M. (2018). Facebook in the News: Social media, journalism, and public responsibility following the 2016 Trending Topics controversy. *Digital Journalism*, *6*(1), 4–20. https://doi.org/10.1080/21670811.2017.1298044

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war. *Foreign Affairs*, *98*(1), 147–155.

Chow, Z. E. (2019). *EVALUATING THE APPROACHES TO SOCIAL MEDIA LIABILITY FOR PROHIBITED SPEECH*. 1–20. https://nyujilp.org/wp-content/uploads/2019/09/NYI406.pdf

Citron, D. K., & Chesney, R. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security National Security. *HeinOnline*. https://scholarship.law.bu.edu/faculty_scholarship/640

CyberPeace Institute. (2021). *Accountability for illegal surveillance by spyware*. https://cyberpeaceinstitute.org/news/accountability-for-illegal-surveillance-by-spyware/

Darras, T. (2011). *Portret van het Nederlandse portretrecht*. https://libstore.ugent.be/fulltxt/RUG01/001/787/253/RUG01-001787253_2012_0001_AC.pdf

de Ruiter, A. (2021). The Distinct Wrong of Deepfakes. *Philosophy and Technology*, *0123456789*. https://doi.org/10.1007/s13347-021-00459-2

Eaton, A. A., Jacobs, H., & Ruvalcaba, Y. (2017). 2017 Nationwide Online Study of Nonconsensual Porn Victimization and Perpetration. *Cyber Civil Rights Initiative*, *June*, 1–28. https://www.cybercivilrights.org/our-services/

European Commission. (2017). *Communication on Tackling Illegal Content Online*. 1–20.

European Commission. (2018a). Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online. *Official Journal of the European Union*, *L 63*(December 2014), 50–61. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2018.063.01.0050.01.ENG&toc=OJ:L:2018:063:TOC

European Commission. (2018b). *REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on preventing the dissemination of terrorist content online*. *0331*(September).

European Commission. (2021). Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. *The EU Artificial Intelligence Act*, *0106*. https://doi.org/10.4324/9781003319436

Europol. (2022). Facing reality ? Law enforcement and the challenge of deepfakes. *Publications Office of the European Union, Luxembourg*. https://doi.org/10.2813/08370

Fernandes, S., Raj, S., Ewetz, R., Pannu, J. S., Kumar Jha, S., Ortiz, E., Vintila, I., & Salter, M. (2020). Detecting deepfake videos using attribution-based confidence metric. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2020-June*, 1250–1259. https://doi.org/10.1109/CVPRW50498.2020.00162

Friedman, L. M., & Hayden, G. M. (2017). Law: Formal and Informal. In *American Law: An Introduction* (pp. 19–34). https://doi.org/10.1093/acprof:oso/9780190460587.003.0002

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning Learning. *Nature*, *29*(7553), 1–73.

Google. (2022). *Beleid tegen misleidende informatie*. https://support.google.com/youtube/answer/10834785?hl=nl

Granot, Y., Balcetis, E., Feigenson, N., & Tyler, T. (2018). In the Eyes of the Law: Perception Versus Reality in Appraisals of Video Evidence. *Psychology, Public Policy, and Law*, *24*(1), 93–104. https://doi.org/10.1097/iop.0000000000000025

GroenLinks. (2020). *Deepnudes verspreiden moet strafbaar worden*. https://groenlinks.nl/nieuws/deepnudes-verspreiden-moet-strafbaar-worden

Guarnera, L., Giudice, O., & Battiato, S. (2020). DeepFake detection by analyzing convolutional traces. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2020-June*, 2841–2850. https://doi.org/10.1109/CVPRW50498.2020.00341

Guera, D., & Delp, E. J. (2019). Deepfake Video Detection Using Recurrent Neural Networks. *Proceedings of AVSS 2018 - 2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance*. https://doi.org/10.1109/AVSS.2018.8639163

Hancock, J. T., & Bailenson, J. N. (2021). The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 149–152. https://doi.org/10.1089/cyber.2021.29208.jth

Heijnen, C. (2021). *Deepfakes en de doeleinden van het Nederlandse aansprakelijkheidsrecht*. *11920041*, 1–43.

Hsu, C. C., Zhuang, Y. X., & Lee, C. Y. (2020). Deep fake image detection based on pairwise learning. *Applied Sciences (Switzerland)*, *10*(1). https://doi.org/10.3390/app10010370

Jones, J. S., Alexander, C., Wynn, B. N., Rossman, L., & Dunnuck, C. (2009). Why Women Don't Report Sexual Assault to the Police: The Influence of Psychosocial Variables and Traumatic Injury. *Journal of Emergency Medicine*, *36*(4), 417–424. https://doi.org/10.1016/j.jemermed.2007.10.077

Jørgensen, R. F., & Zuleta, L. (2020). Private Governance of Freedom of Expression on Social Media Platforms. *Nordicom Review*, *41*(1), 51–67.

Kerner, C., & Risse, M. (2021). Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. *Moral Philosophy and Politics*, *8*(1), 81–108. https://doi.org/10.1515/mopp-2020-0024

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? In *Business Horizons* (Vol. 63, Issue 2, pp. 135–146). https://doi.org/10.1016/j.bushor.2019.11.006

Kietzmann, T. C., Cognition, M. R. C., & Unit, B. S. (2019). Deep Neural Networks in Computational Neuroscience. *Oxford Research Encyclopaedia of Neuroscience*, *January*, 1–29.

Kotsadam, A., & Jakobsson, N. (2011). Do laws affect attitudes? An assessment of the Norwegian prostitution law using longitudinal data. *International Review of Law and Economics*, *31*(2), 103–115. https://doi.org/10.1016/j.irle.2011.03.001

KRO-NCRV. (2020). *Deepfake Therapy*. https://www.2doc.nl/documentaires/series/2doc/kort/2020/deepfake-therapy.html

Lessig, L. (1999). *The law of the horse: What cyberlaw might teach*. https://doi.org/10.1007/bf02763504

Lima, G., Han, J., & Cha, M. (2022). Others Are to Blame: Whom People Consider Responsible for Online Misinformation. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–25. https://doi.org/10.1145/3512953

Maddocks, S. (2020). ‘A Deepfake Porn Plot Intended to Silence Me’: exploring continuities between pornographic and ‘political’ deep fakes. *Porn Studies*, *7*(4), 415–423. https://doi.org/10.1080/23268743.2020.1757499

Meijers Committee. (2022). *CM 2203 Comments on the AI Regulation Proposal*. *June*, 1–8.

Meta. (2022). *Manipulated media*. https://transparency.fb.com/en-gb/policies/community-standards/manipulated-media/

Ministerie van Justitie en Veiligheid. (2020). *Antwoorden Kamervragen over non-consensuele naaktbeelden en pornografie, kinder- en wraakporno op pornosites*. https://www.rijksoverheid.nl/documenten/kamerstukken/2020/09/28/antwoorden-kamervragen-over-non-consensuele-naaktbeelden-en-pornografie-kinder-en-wraakporno-op-pornosites

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions Don’t Lie: An Audio-Visual Deepfake Detection Method using Affective Cues. *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, 2823–2832. https://doi.org/10.1145/3394171.3413570

Ng, E. (2020). No Grand Pronouncements Here..: Reflections on Cancel Culture and Digital Media Participation. *Television and New Media*, *21*(6), 621–627. https://doi.org/10.1177/1527476420918828

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, C. M., Nguyen, D., Nguyen, D. T., & Nahavandi, S. (2021). *Deep Learning for Deepfakes Creation and Detection: A Survey*. 1–16. http://arxiv.org/abs/1909.11573

NOS. (2017). *Pechtold vindt nepfoto die Wilders stuurde onacceptabel*. https://nos.nl/artikel/2156807-pechtold-vindt-nepfoto-die-wilders-stuurde-onacceptabel

O’Donnell, N. (2021). Have we no decency? section 230 and the liability of social media companies for deepfake videos. *University of Illinois Law Review*, *2021*(3), 701–740.

Oktay, D., & Bala, H. A. (2015). A holistic research approach to measuring urban identity: Findings from Girne (Kyrenia) area study. *Archnet-IJAR*, *9*(2), 201–215. https://doi.org/10.26687/archnet-ijar.v9i2.687

Openbaar Ministerie. (2020). *Strafvorderingsrichtlijn misbruik seksueel beeldmateriaal*. https://www.om.nl/actueel/nieuws/2020/12/03/strafvorderingsrichtlijn-misbruik-seksueel-beeldmateriaal

Porter, G., & Kennedy, M. (2012). Photographic truth and evidence. *Australian Journal of Forensic Sciences*, *44*(2), 183–192.

Reddit. (2020). *Updates to Our Policy Around Impersonation*. https://www.reddit.com/r/redditsecurity/comments/emd7yx/updates_to_our_policy_around_impersonation/

Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European Journal of Operational Research*, *291*(3), 906–917. https://doi.org/10.1016/j.ejor.2020.09.020

Riis, T., & Schwemer, S. F. (2019). Leaving the European Safe Harbor, Sailing Towards Algorithmic Content Regulation. *SSRN Electronic Journal*, *22*(2019). https://doi.org/10.2139/ssrn.3300159

Rijksoverheid. (2022). *Hoe hoog zijn de boetes in Nederland?* https://www.rijksoverheid.nl/onderwerpen/straffen-en-maatregelen/vraag-en-antwoord/hoe-hoog-zijn-de-boetes-in-nederland

RTL Nieuws. (2021, April 24). *Tweede Kamer had videogesprek met deepfake-imitatie van stafchef Navalny*. https://www.rtlnieuws.nl/nieuws/nederland/artikel/5227214/tweede-kamer-navalny-deepfake-volkov

Savin, A. (2021). The EU Digital Services Act: Towards a More Responsible Internet. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Schermer, B., & Sloot, B. van der. (2020). *Het recht op privacy in horizontale verhoudingen*. https://repository.wodc.nl/bitstream/handle/20.500.12832/2472/3062_Volledige_Tekst_tcm28-457676.pdf?sequence=2&isAllowed=y

Somers, M. (2020). *Deepfakes, explained*. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

Spencer, C., Mallory, A., Toews, M., Stith, S., & Wood, L. (2017). Why Sexual Assault Survivors Do Not Report to Universities: A Feminist Analysis. *Family Relations*, *66*(1), 166–179. https://doi.org/10.1111/fare.12241

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Transactions on Graphics*, *36*(4). https://doi.org/10.1145/3072959.3073640

Tariq, S., Lee, S., & Woo, S. (2021). One detector to rule them all: Towards a general deepfake attack detection framework. *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*, 3625–3637. https://doi.org/10.1145/3442381.3449809

The New York Times. (2018). *Truth in a Post-Truth Era: Sandy Hook Families Sue Alex Jones, Conspiracy Theorist*. https://www-media.floridabar.org/uploads/2018/09/18-RW-Sandy-Hook-Families-Sue-Alex-Jones.pdf

TikTok. (2020). *Combating misinformation and election interference on TikTok*. https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok

Tweede Kamer der Staten-Generaal. (2019). *Wijziging van onder meer het Wetboek van Strafrecht in verband met de herwaardering van de strafbaarstelling van enkele actuele delictsvormen (herwaardering strafbaarstelling actuele delictsvormen)*. *3*, 1–26. https://zoek.officielebekendmakingen.nl/kst-35080-7.pdf

Volpe, A. (2018). Deepfake Porn Has Terrifying Implications. But What If It Could Be Used for Good? *Men's Health*. https://www.menshealth.com/sex-women/a19755663/deepfakes-porn-reddit-pornhub/

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11), 39–52. https://doi.org/10.22215/TIMREVIEW/1282

Woodley, X. M., & Lockard, M. (2016). Womanism and snowball sampling: Engaging marginalized populations in holistic research. *Qualitative Report*, *21*(2), 321–329. https://doi.org/10.46743/2160-3715/2016.2198

Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality*, *11*(3). https://doi.org/10.1145/3309699

Zhao, F., Fashola, O. I., Olarewaju, T. I., & Onwumere, I. (2021). Smart city research: A holistic and state-of-the-art literature review. *Cities*, *119*(May 2020), 103406. https://doi.org/10.1016/j.cities.2021.103406

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., & Xia, W. (2020). *Learning Self-Consistency for Deepfake Detection*. 15023–15033. http://arxiv.org/abs/2012.09311

Jurisprudence

ECLI:NL:RBAMS:2005:AS4748

ECLI:NL:RBDHA:2021:1885

ECLI:NL:RBLIM:2014:9244

ECLI:NL:RBDHA:2019:11523

ECLI:NL:RBMNE:2020:24

ECLI:NL:RBAMS:2019:8415

ECLI:NL:RBAMS:2019:7357

ECLI:NL:RBAMS:2012:BX7043