TILBURG UNIVERSITY

Understanding Society

Tilburg Papers in Culture Studies

Paper 237

Deep fakes – an emerging risk to individuals and societies alike

by

Tünde Faragó©

(Tilburg University)

December 2019

Deep fakes – an emerging risk to individuals and societies alike

MA Thesis

Name of author: Tünde Faragó

Student number: 2031520

MA track: Online Culture

Major: Art, Media and Society

Department of Culture Studies

School of Humanities and Digital Sciences

Date: August 2019

Supervisor: dr. Piia Varis

Second reader: prof. dr. Jan Blommaert

Contents

1. Introduction 3
2. Framework 6
2.1. Post-truth 6
2.2. Deep fakes and moral panic 12
3. Methodology 18
4. Case study analyses 19
4.1. Deep fakes and their potential to harm individuals 19
4.2. Deep fakes and their potential to harm societies 30
5. Conclusion 44
6. References 53

1. Introduction

Digitalization and the rise of the internet have brought about a revolution in the way people can access information and share it with others. It also resulted in citizens themselves becoming producers of knowledge, a privilege that once more strictly belonged only to governments, media outlets and the academia. (Hanell and Salö, 2017: pp. 155-156) But with traditional institutions losing authority over facts, fake news, misinformation and lies have become a commonplace, sidelining the truth in the process. (Llorente, 2017: p. 9)

Although fake news, propaganda and lies are tales as old as time, it wasn't until the year 2016 when major online misinformation campaigns that seem to have contributed to Donald Trump becoming the president of the United States and British people voting to leave the European Union, that these phenomena became a center of attention in Western societies. These, for many shocking events, resulted in Oxford Dictionaries to name "post-truth" the word of the year in 2016. (Oxford Dictionaries, n.d.)

This post-truth era that we live in now is said to be guided by our emotions as objective facts and the truth are cherry-picked and shaped to fit our own versions of reality. (McIntyre, 2018: chapter 1, paragraph 12) But humanity's ability to twist the reality has taken an even a bigger step forward in recent years with the creation of deep fake technology. With this technology one can create deep fakes: videos of real people saying and doing things they have never said or done in real life. (Chesney and Citron, 2018: n.p.) In 2017 the first deep fakes emerged in which faces of young Hollywood actresses, such as Scarlet Johansson and others, were superimposed into porn movies. The technology that was once only in the hands of experts was soon democratized and web pages and free apps promising to create video forgeries of celebrities, ex-girlfriends, colleagues and even strangers popped up all around the world. This has led to a situation where suddenly everybody could fall victim to a deep fake. (Schwartz, 2018) Moreover, the machine learning ability of Artificial Intelligence, which was used to create deep fakes, is developing so fast that it becomes very hard for software, not to mention are very own human senses, to detect the forgeries. (Chesney and Citron, 2018: n.p.) And exactly this inability of people to tell them apart from reality is where their true danger lies.

Since deep fakes are a relatively new thing, besides the report done by Law professors Bobby Chesney and Danielle Citron in 2018 about the looming challenges of deep fakes, there have not been any other major studies dealing with this phenomenon. (Chesney and Citron, 2018) Therefore, this thesis represents one of the first attempts to explore their potential social, cultural and political implications through the analyses of a few carefully selected case studies. The aim of this thesis is to analyze the social, cultural and political ramifications of deep fakes in relation to individuals and societies alike.

The first two chapters provide a theoretical background, analyses of current societal developments and explanation of the deep fake technology. In the third chapter an explanation will be given about the methodology chosen to analyze the phenomenon of deep fakes. The following chapters will provide an analysis of the cases utilizing the aforementioned framework. The case studies chosen for this thesis represent examples where deep fakes had some serious consequences on individuals and societies. For the chapter about the potential of deep fakes to harm individuals, two recent case studies were chosen where the victims suffered psychological and reputational damages caused by forged videos. The first case study from 2018 is about a young Muslim freelance journalist from India, Rana Ayyub, who was harassed and bullied by a pornographic deep fake intended to silence her activism. The second case is from 2012 and focuses on an Australian woman, called Noelle Martin, whose pictures were doctored into adult movies and sexually explicit images and uploaded onto famous porn websites. These particular cases were chosen because there is plenty of information available about them and at the same time, they provide good insights to the various consequences deep fakes might have on individuals. On the other hand, to analyze what kind of potential risks deep fakes pose for societies, a Belgian deep fake example was chosen where a political party commissioned a deep fake of Donald Trump in order to shape the public's opinion about a political matter. Although this specific case was only small-scale, nevertheless it showed the potential capacity of such technology to cause some serious disruption in a society. Since there has not been any other major incident which involves deep fakes, other further examples of doctored videos, fake news and misinformation campaigns are also used to explain as to what we might expect when it comes to deep fakes. The final chapter summarizes the central points of the analyses of the above-mentioned case studies and at the same time provides some suggestions for future studies in this field.

2. Framework

2.1.Post-truth

In March 2018 following the horrific Parkland high school shooting in the U.S., a GIF of one of the shooting survivors, Emma González, tearing up a copy of the U.S. Constitution started circulating on social media. Even though the animation was fake it still stirred up a lot of negative emotions, especially among conservative supporters. (Mezzofiore, 2018)

González first became known to the public as she gave a passionate speech followed by a prolonged moment of silence at a gun control rally in Washington only four days after a gunman killed 17 people at her school on 14th of February 2018. In just a few weeks she became one of the faces of the campaign #NeverAgain which attracted a lot of support all across America. (Mikkelson, 2018; Mezzofiore, 2018) However, just hours after the Constitution ripping GIF went viral, many that previously had supported her now turned against her. Shortly after the fake GIF appeared on the internet, multiple sources pointed out that it was doctored; however, it did not stop it from spreading on social media platforms and websites like a wildfire. The original video depicted González ripping in two a shooting target as part of the Teen Vogue story about the survivors released just days before the fake GIF started circulating the internet. (Mezzofiore, 2018)

Soon after the fake GIF appeared, the shooting survivor and vocal activist quickly became a target of social media attacks, many calling her a traitor. Jesse Hughes, the frontman of the band Eagles of Death Metal, was among the first who accused González in a Tweet for treason. (Snapes, 2018) Prominent right-wing figures, such as the famous actor, Adam Baldwin, who is known for his conservative comments, immediately jumped to the chance to share the image on Twitter. Although he later deleted his Twitter post he still defended the fake GIF as “a political satire”. (Mezzofiore, 2018)

The moral of this story is clear: the truth becomes irrelevant in the heat of the moment while feelings and opinions dictate the perspective on reality. (Llorente, 2017: p. 9) As "[e]xperts and facts no longer seem capable of settling arguments to the extent that they once did" (Davies, 2018: p. xiv), we now live in what many call the post-truth era, "where objectivity and rationality give way to emotions, or to a willingness to uphold beliefs even though the facts show otherwise." (Llorente, 2017: p. 9)

In 2016 Oxford Dictionary named "post-truth" as the word of the year following controversial events and surprises on the political stage like Donald Trump winning the U.S. presidential elections or Britain's decision to leave the EU. (Llorente, 2017: p. 9) They defined this term as "[r]elating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief." (Oxford Dictionaries, n.d.) However, this does not mean that we are *past* the truth or that before we had more *truthful* periods; what this definition refers to is that we live in an era that is indifferent to the truth. This is not to say that people do not care about the truth anymore; it only means that the truth becomes a matter of opinion. Whatever is in line with people's personal beliefs will more likely be considered to be true, regardless of the objective facts.

This is the era where fake news and doctored GIFs have become commonplace, sidelining the objective facts and experts. And very recently fake videos, called deep fakes, emerged promising new ways in which reality can be bent and altered to favor some opinions over others. In the light of all this,

> "[t]he value or credibility of the media has somewhat faded in comparison to personal opinions. The facts themselves take second place, while 'how' a story is told takes precedence over 'what'. It is therefore not about what has happened, but rather about listening, seeing and reading the version of facts which more closely fits with each person's ideology." (Llorente, 2017: p. 9)

According to Davies (2018: p. 7), appealing to people with objective facts or expert opinions rarely triggers them to act in a certain way. Instead, people are easily swept away if one plays on their emotions. Therefore, images or stories which change the way people feel will be exchanged very quickly among people who share the same sentiments. A public that is susceptible to emotional appeals is stirred and shaped quite easily by fake news, rumors or lies. However, we need to be careful when we talk about the post-truth and lies. These two may be connected but they are not the same; post-truth can be more accurately understood as the relativization of the truth. (Zarzalejos, 2017: p. 11)

According to McIntyre the word post-truth is "irreducibly normative" (McIntyre, 2018: chapter 1, paragraph 6). "[T]here are many ways one can fit underneath the post-truth umbrella." (McIntyre, 2018: chapter 1, paragraph 11) One can be willfully indifferent to the truth, not knowingly spread a lie, or even intentionally lie and falsify something. (*Ibid*: chapter 1, paragraph 8-10) However, this is not something new. What is new about the post-truth era is the "existence of reality itself". (*Ibid*: chapter 1, paragraph 11) The reason for this is simple: if the majority of society or its leaders are in dispute over the basic facts, our whole democratic systems are at risk. (*Ibid*)

As McIntyre (2018: chapter 1, paragraph 12) points out, "the real problem here…is not merely the content of any particular (outrageous) belief, but the overarching idea that – depending on what one wants to be true – some facts matter more than others." This means that for many people what corresponds with their belief and opinions is true. The facts are in this sense not disregarded; they are just cherry-picked and shaped to fit a person's belief about reality. And this is at the core of the problem, since it directly "undermines the idea that *some things are true irrespective of how we feel about them*." (*Ibid*, emphasis original) For instance the idea that the Earth is not flat, or that our planet revolves around the Sun: these scientific facts will still be true even if conspiracy theorists do not believe in them. However, when the majority of the population believes in conspiracy theories or fake news then it becomes a serious threat to the shared understanding of the nature of reality.

We thus live in a world which “revolves around passions and beliefs; where truth is no longer needed” (Madeiros, 2017: p. 24), because each person has his or her own version of the truth that corresponds to their personal ideology. And when facts are shaped and dictated by one’s political ideology then we have a clear recipe for political dominance. (McIntyre, 2018: chapter 1, paragraph 13)

This is also not a new phenomenon as such. What is post-truth today was called propaganda in other times. (Medrán, 2017: p. 33) People who participated in the power structures of a society were able to produce and legitimize certain types of knowledge and introduce them as credible. (Hanell and Salö, 2017: p. 155) “Knowledge, thus, is a stratified construct, interest-laden and historically saturated with power, and…some people have more power to speak about some subjects than others.” (*Ibid*) According to this, not all types of knowledge are equally credible or equally visible for that matter. However, what is different nowadays is that everybody can have access to all kinds of different sources of information, create their own facts and share them online. With the rise of the internet, the authority over the production of knowledge and its legitimization no longer is in the hands of the state, the academia or traditional media. Considering the fact, that nowadays anyone can produce and share ideas online, knowledge of different orders of visibility is intertwined on the internet. (*Ibid*: pp. 155-156)

> “Today, access to informative content, as well as its immediacy and volume, has no precedents. The impact of digitalization in the world of communications has brought about a revolution in a way that people can produce information themselves…Similarly, it has changed the way they consume and assimilate it.” (Gooch, 2017: p. 14)

At the same time, while everybody can claim that what they know is true, the trust in expert opinion has drastically declined. "Journalists, judges, experts and various other 'elites' are under fire today." (Davies, 2018: p. 26) This is what happens when the world gets taken over by feelings. (*Ibid*: pp. xvi-xvii)

While digitalization allowed people to access, produce and share all kinds of information online, it has also contributed to the creation of filter bubbles: places "where we find only the news we expect and the political perspectives we already hold dear." (Gillespie, 2014: p. 188) Chesney and Citron (2018: p. 13) argue that even without technology, people tend to surround themselves with ideas that go hand in hand with their own personal beliefs. Social media platforms further exacerbate this tendency by allowing their users to create and share content to their own liking. (*Ibid*) Moreover,

> "[p]latforms' algorithms highlight popular information, especially if it has been shared by friends, surrounding us with content from relatively homogenous groups. As endorsements and shares accumulate, the chances for an algorithmic boost increases. After seeing friends' recommendations online, individuals tend to share them with their networks." (*Ibid*)

Thus, information that is shared more often by users becomes more visible and at the same time is considered more credible.

What I already mentioned above regarding the status of facts or the truth in the 'post-truth' era also nowadays goes for instance politics. That is, it does not necessarily mean that the facts or the truth are completely disregarded, "but instead a conviction that facts can always be shaded, selected, and presented within a political context that favors one interpretation of truth over another." (McIntyre, 2018: chapter 1, paragraph 4) This is very evident when we take a look at the Trump administration, which is regarded as the first ever

Twitter-based presidency in the world. (Hollinger, 2017) Trump and his associates very often play with facts or “alternative facts”, as they sometimes refer to it, to present their own version of events. (McIntyre, 2018: chapter 1, paragraph 4) For instance, back in January 2017 Trump’s team defended his press secretary’s comments about how many people attended the inauguration, saying that the secretary did not lie but offered “alternative facts”. (Swaine, 2017) It seems that

> “[i]t doesn't matter if [Trump’s] comments are true -- and multiple fact-checking sites like PolitiFact, FactCheck.org and the Washington Post's Fact Checker blog have shown that many of the assertions he tweets are false. Trump's 140-character outbursts are just what many among his 41.5 million online followers want to hear.” (Collins, 2018)

However, this atmosphere is not only U.S.-specific; it is present all over the world. Traditional media is benched while alternative media is seen as more trustworthy. This is no surprise as, as it was said before, traditional media and experts have lost their credibility in the public’s eye. And the U.S. president calling certain media houses the creators of “fake news” and “the true enemy of the people” just adds additional fuel to the fire. (Bell, 2019; Wagner, 2018) In such a climate people will turn to someone they can trust the most – to themselves and their own convictions. It is not surprising perhaps that they stop questioning the credibility of the facts presented to them. After all, everybody thinks of themselves and their friends as trustworthy sources. (Sundar, 2016)

This makes people very vulnerable to fake news, lies and propaganda. In the absence of an “institution to establish filters, separate the wheat from the chaff and put different views into perspective” all we are left with is our own beliefs and opinions to guide us. (Madeiros, 2017: p. 23) Very often this means that people will try to fit the world into their own perspective and not the other way around.

In hearing what they want to hear, what moves people more than anything is a good juicy story. Objective facts and well researched topics may fall on deaf ears, but a rumor or a hoax that has an emotional appeal and confirms existing biases will be shared among like-minded people. And if a famous person with a bigger social media visibility circulates such a story, like Trump or Adam Baldwin, it will be shared instantly among millions of followers, increasing the story's credibility as well in the process. (Sundar, 2016)

However, it is not that we are so gullible as to believe everything we are presented with; we are rather just incapable of admitting that we are wrong. When people are confronted with a truth that contradicts their beliefs it creates a psychological tension. If a person is confronted with evidence that proves his or her belief is wrong, the logical reaction would be to change one's mistaken belief. But this is not what happens in reality. People are reluctant to let go of their own confirmational biases and very often when confronted with contradicting evidence they are not thinking rationally. (McIntyre, 2018: chapter 3, paragraph 2-3) In politics for instance this plays out by picking a side and sticking to it even if the facts say otherwise. Such a political climate where personal truths matter more than objective ones creates a danger to a society as a whole. How does an emotionally susceptible society even stand a chance against deep fakes: fake videos that are created to look real but are actually not?

2.2.Deep fakes and moral panic

Until recently the technology of deep fakes was limited mostly to the community developing Artificial Intelligence; however, in 2017 a Reddit user known by his username "Deepfakes" developed the same technology with the help of Google's free open source machine learning software called TensorFlow. He used his creation to put celebrity faces onto the bodies of actresses in porn movies. A number of media outlets reported about these fake porn movies, calling them "deepfakes". Although Reddit banned the videos for their violation of privacy and involuntary pornography, the damage was already done. The creator of the deep fakes soon released a free application called FakeApp which allowed anyone who had access to the internet and photos of people to create deep fakes for all kinds of uses. (Schwartz, 2018)

Programs and apps with the ability to alter and tweak images, audios and videos are not a new phenomenon as such. In our daily lives, many of us use filters to smooth out the imperfections of pictures or videos we intend to post to our social media accounts. Some of these tools are quite impressive, yet people can usually tell if an image of a person is altered or not. And what our human eyes miss, forensic technologies are still able to detect. However, deep fake technology is expected to revolutionize all this. Heavily relying on AI technologies, deep fakes use neural networks and more recently GAN technologies (generative adversarial networks) to create realistic forgeries that are nearly impossible to detect. These networks work by mimicking the human brain: “[m]uch as experience refines the brain’s neural nodes, examples train the neural network system.” (Chesney and Citron, 2018: p. 5) With enough examples the neural networks are able to recreate very accurate impersonations. The GAN technology is even more sophisticated. It was created by Ian Goodfellow, a researcher at Google. (*Ibid*)

> “A GAN consists of two neural networks playing a game with each other.
> The *discriminator* tries to determine whether information is real or fake. The other neural network, called a *generator*, tries to create data that the discriminator thinks is real.” (Hergott, 2019)

The discriminator is able to very accurately determine fakes but given enough time and practice, the generator can create forgeries that will fool even the discriminator network, not to mention our human senses.

Since the 2017 deep fake Reddit incident this mind-bending technology has become widely available to everyone and many deep fake websites and apps have emerged, offering to create forged videos of celebrities, ex-girlfriends or anyone you desire for as little as 20 dollars or sometimes even free of charge. (Harwell, 2018) All you need to provide is a few hundred photos of a person or a short video. (Solsman, 2019a) Nowadays it is of course also extremely easy to find photos of people online, as after all we live our lives mostly in the public eye, by posting regularly on social media platforms. Also, with the rapidly advancing AI technologies, now even just a handful of pictures will be enough to make forgeries. The deep fake technology developed by Samsung researchers in 2019 enables its users to create video impersonations of anyone based on just a single picture. (Solsman, 2019b) And since this technology requires only one image of a person, people who do not necessarily have so many pictures of themselves online are also at risk of becoming targets of deep fake manipulations.

Up until now, the public perception of deep fakes has been mostly negative. News media outlets and experts supercharge this negativity with articles and reports warning about their potential to create nightmarish scenarios for both individuals and governments. (Chesney and Citron, 2018; Harwell, 2018; Schwartz, 2018) As boyd (2012) argues, "[m]oral panics emerge whenever something new happens that disrupts the social order in a way that makes people anxious and afraid." There is a certain amount of moral panic every time new technology appears (*Ibid*), and deep fakes are no exception to this. However, this is not to say that deep fake technology is inherently bad, far from it. According to Kranzberg's law we should not label any technology as good or bad or even neutral: "[a] technology's value is shaped by its social construction—how designers create it and how people use it, interpret it, and reconfigure it. It is not an outcome of the technology alone or its potential." (boyd, 2008: p. 12) How a certain technology is regarded or valued very much depends on societal developments. A doctored video two decades ago, when YouTube was in its early stages, would have been regarded as a fun prank, but today when people are unable to distinguish between what is real and what is fake, a decent deep fake can pass as a real thing. And it is not only because the technology is so advanced that forgeries can easily fool people. Nowadays, even a crudely doctored video will do that. For instance, a poorly altered video of U.S. House Speaker Nancy Pelosi to make her sound drunk gained more than 2 million views on the first day it was posted online.[1] (Harwell, 2019) In this media era, fake news can very quickly become widely known and accepted, due to algorithmic boosting and filter bubbles which help spread this kind of content extremely fast. The Nancy Pelosi case also shows that the threat of deep fakes is real, especially in this post-truth era where people's emotions take precedent over the truth itself. That is, people rely on their emotions and personal beliefs rather than the real facts when deciding what is true and what not.

In an environment like this, a good fake story that plays on people's emotions is in fact a very powerful tool that could lead an armed man onto a crusade to save children from an underground sex-trafficking ring from a Washington pizza place. Back in 2016, amid the U.S. presidential elections campaigns, WikiLeaks published hacked emails by Hillary Clinton and her staff members which sparked a whole range of conspiracy theories. The hacked emails among other things contained the word "pizza" and dinner plans between Clinton's top aides. It did not take long for Trump enthusiasts, who were desperately looking for some clues of wrongdoings in the leaked documents, to create a connection between the phrase "cheese pizza" and child pornography. This connection soon led the conspiracy theorists to believe that Hillary Clinton and her administration are running an underground child-trafficking ring from a Washington pizza parlor called Comet Ping Pong. (Aisch, Huang and

[1] The doctored video of Nancy Pelosi can be accessed through this link: https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/?noredirect=on&utm_term=.66ae49201634

Kang, 2016) “The theory started snowballing, taking on the meme #PizzaGate. Fake news articles emerged and were spread on Twitter and Facebook.” (*Ibid*) The fake stories ranged from kill rooms and Satanism all the way to theories about cannibalism taking place in the supposed underground tunnels of the pizza parlor. This is when a 28-year-old North Carolina man, called Edgar M. Welch, decided to take matters into his own hands and save the day. Armed with military rifles and other guns he arrived at the Comet on 28th of December 2016 and fired three warning shots before surrendering to the police after having found zero evidence to support his theories of sex-trafficking and pedophilia. However, this did not put an end to the conspiracy theories. On the contrary, supporters of the theories refused to let go of their personal beliefs and went so far as to accuse the mainstream media of helping to cover up the secret crime organization. (*Ibid*)

We can only imagine what even a low-tech deep fake might do in such a climate where only a handful of emails about “cheese pizza” managed to conjure up a conspiracy theory that is accepted by thousands as true. Again, this is not because people are so gullible or naïve as to believe whatever you serve to them. Instead, fake news and conspiracy theories work on people, because they are willing to accept all kinds of (mis)information that fits with their own world views. This environment is what Davies calls a nervous state, “a state of constant and heightened alertness” of individuals and governments alike. (Davies, 2018: p. xii) He describes this as a state where people often make decisions in a heat of the moment, while rational thinking is benched, and emotions are the ones stirring the wheel. (*Ibid*)

Social media further aggravates this nervous environment. (Davies, 2018: p. 6) On the internet “information moves like a virus through a network”, including misinformation. (*Ibid*) In such a climate images and words are used as tools to engage and mobilize people, very often ignoring the validity of the information they provide or its correspondence with the objective reality. (*Ibid*: p. 7) “This is the anxiety that now surrounds ‘fake news’ and propaganda." (*Ibid*) It is this nervous state that also made people believe all kinds of hoaxes and fake news following a deadly tsunami that hit Indonesia in 2018. While recovering from the devastating earthquake and tsunami that hit the Indonesian Sulawesi island in September 2018 doctored images, videos and fake news alleging that another much deadlier earthquake and volcanic eruptions will occur in the next few days started spreading online and in many places where the electricity was cut due to the severe weather, by word of the mouth. The rumors caused major public panic and shock. The situation escalated so much that the government had to release official statements urging the public not to fall for such misinformation and to report it immediately to the police. (Lyons and Lamb, 2018; The Star Online, 2018) While things returned to normalcy since the government took extra measures to debunk the fakes, "Indonesia's communications ministry has announced plans to hold weekly briefings on fake news, in an effort to educate the public about the spread of disinformation." (Lamb, 2018)

Nowadays, this public anxiety is indeed much heightened all around the world. And this creates an excellent breeding ground for misinformation, fake news and most recently deep fakes. Although deep fakes are a relatively new phenomenon, experts already warn about the risks they can pose to individuals, organizations or to entire countries. Nobody is safe from deep fakes. Our society already "*suffers from truth decay as our networked information environment interacts in toxic ways with our cognitive biases*." (Chesney and Citron, 2018: n.p., emphasis original) Add to this the new technology of deep fakes and what we get is new ways in which people, governments and democratic institutions can be exploited, sabotaged and altogether endangered. Therefore, in this thesis, the social, cultural and political ramifications of deep fakes will be analyzed in relation to individuals and societies alike.

3. Methodology

In this thesis a case study method will be used to analyze the social, cultural and political ramifications of deep fakes in connection to individuals and societies. This method is a very popular choice in social sciences especially “when a holistic, in-depth investigation is required.” (Zainal, 2007: p. 1) Case studies are designed to provide an in-depth analysis of a contemporary phenomenon in its own context and environment through a limited amount of cases. (*Ibid*: pp. 1-2) In other words a case study is “an empirical inquiry that investigates a contemporary phenomenon within its real-life context; when the boundaries between phenomenon and context are not clearly evident; and in which multiple sources of evidence are used.” (Yin, 1984 as cited in Zainal, 2007: p. 2) Since deep fakes are a relatively new phenomenon and constantly evolving, a case study method seemed as the best choice, as it offered a unique possibility of analyzing a small number of cases in great detail on a micro level. The lack of relevant research about deep fakes presents yet another reason why this approach was chosen. Moreover, this method provided an opportunity for multiple sources of data regarding specific deep fake cases to be explored and investigated, giving insights to a previously less known phenomenon.

As deep fake technology is rapidly evolving, there are many cases to be explored in connections to this technology. However, given the limited scope of this thesis I had to draw the line somewhere. I am aware that this limits my research in the sense that I am unable to provide general conclusions based on the limited number of cases analyzed. Nonetheless, the specific cases in this thesis were chosen so as to be able to explore different social, cultural and political aspects of deep fakes and to analyze the potential risks they pose to both individuals and societies. In the upcoming chapters, the cases selected for this thesis will be discussed.

4. Case study analyses

4.1.Deep fakes and their potential to harm individuals

People have of course lied to, cheated and framed other people since the beginning of time, and

> “[a]ll of this will be true with deep fakes as well, only more so due to their inherent credibility and the manner in which they hide the liar’s creative role. Deep fakes will emerge as powerful mechanisms for some to exploit and sabotage others.” (Chesney and Citron, 2018: p. 16)

The specifics of these mechanisms in using deep fakes to harm other people are yet to be largely seen. Possible uses include just a joke to embarrass a colleague, a porn video for someone’s own gratification, identity theft or even to spur violence. (Chesney and Citron, 2018: p. 17) Only the sky is the limit for the many uses of deep fakes, or in other words the infinite creativity of one’s imagination.

Although nobody is safe from deep fakes, some people are more vulnerable than others. In December 2018 Scarlet Johansson, one of the celebrity victims of pornographic deep fakes, stated that although these fakes are demeaning, they do not necessarily affect her, because people mostly know it is not her in the porn videos. But while her celebrity status helps her battle these fakes much easier, she stresses that it is a totally different story for unknown people who can lose their jobs over their image being doctored to fake videos. (Gordon, 2019) However, in some cases even being publicly known will not save one from intentional reputational sabotage.

Some say Rana Ayyub is one of the most abused women in India in recent history. Besides being a well-known investigative female journalist in a country where the gender gap is still very high[2], she is also Muslim and is even considered to be part of the anti-establishment. (Ayyub, 2018; World Economic Forum, 2018) As she says of herself, she "ticks all the boxes." (Ayyub, 2018) Yet despite years of constant trolling and hatred on social media which was intended to silence her, she kept reporting anyway about sensitive cases regarding human rights violations. These cases were usually involving high-profile individuals very often in governmental positions. Although over the years she managed to ignore and endure most of the online violence and hatred targeted towards her, in April 2018 all that changed. After a horrific sexual assault on an 8-year-old girl in Kashmir in 2018, Ayyub was invited by the BBC and Al Jazeera to comment on India's stance towards child molesting and the country's shameful behavior in protecting the perpetrators. This sparked an outrage among the supporters of the ruling Hindu nationalist party, Bharatiya Janata Party (BJP in short), which decided to teach her a lesson. In the upcoming days, series of offensive fake tweets appeared claiming to be written by Ayyub. Among others they stated hatred towards India and the people of India. (*Ibid*) One even read "I love child rapists and if they are doing it in the name of Islam I support them." (as cited by Ayyub, 2018) The fake tweets were shared by thousands and needless to say, soon her social media was flooded with abuse. Eventually, Ayyub released an official statement that she was not the author of the alleged tweets; however, the abuse did not stop there. Not long after, a deep fake emerged with Ayyub's face doctored onto a young porn actress's body. The BJP fans shared the fake video on their official website and things escalated really quickly from there onwards. Instantaneously it was viewed and shared by thousands of people, many part of India's political establishment. Her private contact information was also published by the unknown

[2] In 2018 India ranked 108th among 149 countries on the gender equality scale according to the World Economic Forum report from 2018.

perpetrators and the inappropriate offers for sex soon started arriving on her private phone and social media accounts. (Ayyub, 2018)

It goes without saying that she was beyond devastated. In just a mere two days her entire public reputation was destroyed. People were calling her and texting her to ask if the video was real, while others were shaming her online, referring to her as a Jihadist who favors child rapists, some even called for her deportation to Pakistan. To make the matter even worse, when she tried to file a complaint about the deep fake video to the police, the officers refused to report it. (Ayyub, 2018) She remembers her struggles with the police like this:

> "I couldn't believe it. I was a woman standing in front of them who had mustered up the courage to file a complaint and they were trying to dodge it. I threatened them. I told them if they didn't want to register a complaint then I would write about them on social media. Finally, after my lawyer told them we would go to the media, they filed the report." (Ayyub, 2018)

The coordinated misinformation campaign to smear her name seemed to work just as it was intended. People's own convictions and confirmational biases were determining Ayyub's public narrative. Supporters of the establishment and even the authorities shaped their opinion about her even before reviewing the evidence. Therefore, it is no surprise that the complaint she filed with the police fell on deaf ears. For months Ayyub did not hear back from them at all. Things only started changing when finally, the United Nations intervened. "Sixteen special rapporteurs wrote to the Indian government" requesting from them to protect Ayyub. (Ayyub, 2018) David Kaye, one of the UN special rapporteurs on freedom of opinion and expression, stated that the UN was especially worried about Ayyub's safety, since the verbal abuses to her reporting went beyond just disliking her writing and turned "sexual and violent in nature". (O'Grady, 2018) He added that the UN feared for Ayyub's life due to the fact that her journalist colleague and friend Gauri Lankesh was killed in front of her house in Bangalore, India by a helmeted biker in September 2017, just a year before the highly coordinated online attacks started targeting Ayyub. Lankesh was just like Ayyub, an outspoken critic of the Hindu nationalist government and its Prime Minister Modi. In her last editorial before she was killed, she alleged "that spreading fake news had contributed to the success of Mr. Modi and his party." (Mondal, 2017) However, she was not the only victim. In the last decade many other journalists who attempted to criticize the current Indian political establishment were also silenced in similar ways as Lankesh. Yet these killings and assaults are not only present in India. Nowadays being a journalist or an outspoken intellectual anywhere in the world is a dangerous profession. According to a report from 2017 "[m]ore than 800 journalists, media workers and social media producers have been killed during the past ten years." (Carlsson and Pöyhtäri, 2017: p. 12) There are however, many more non-lethal attacks that can bring harm to a person ranging from "intimidation, harassment and arbitrary detention to misogynistic attacks directed against women journalists" and very recently deep fakes as well. (*Ibid*) What these statistics indicate is that experts are not just losing the trust of the public, but they are also specifically targeted if they attempt to question the popular narratives, norms and values in a society.

Not long after the UN filed reports to the Indian government, the online and offline abuses against Ayyub decreased rapidly, but the damage was already done. As Ayyub recollects, she is not the same person anymore after this highly coordinated misinformation campaign against her. She became less present online and more cautious of expressing her opinion on social media. (Ayyub, 2018) What we see here is a clear retreat of a victim. Risks and threats to a person's digital security can affect the ways in which one will use technology in the future. (Singh and Roseen, 2018) Ayyub for instance decided to limit her online presence and be less outspoken on social media. (Ayyub, 2018) This just proves how our online world is in fact reinforcing the embedded norms and values of our society. Groups which are already marginalized in our physical world have a higher chance of being victimized online as well. People, like Ayyub, "whose intersecting identities cover multiple marginalized groups face among the worst forms of online abuse." (Singh and Roseen, 2018) Being a Muslim woman and a freelance journalist not afraid to speak out against the government puts a direct target on her back. "And at a time when the technology landscape is evolving in increasingly intrusive and unpredictable ways, these threats can often translate into offline harms, too, as the lines between our digital and real lives are continuously blurred." (Singh and Roseen, 2018) While Ayyub decided to only limit her voice on social media platforms, other victims might choose to disappear from the online world altogether for fear of being targeted and abused again. "As a result, when thinking about the collective outcome that silencing individual voices has in a democratic society, the effects become particularly worrisome." (Singh and Roseen, 2018)

In this environment women are especially vulnerable. Pornographic deep fakes intended to silence, intimidate or exploit women are becoming a new online phenomenon all over the world and not just in India. Back in 2012 when a then 18-year-old Australian woman, named Noelle Martin, googled her own pictures she was horrified with the results. Her pictures taken from her private Facebook account were superimposed onto pornographic images and videos and shared multiple times on famous porn websites and forums. Her face was also used as a cover for porn DVD's. But it is when she tried to speak out about the fake images and videos when the real harassment started. She was doxxed by harassers, who published her private address and contact information in order to terrorize and silence her. Some porn sites and forums even refused to take down the deep fakes, unless she will send them nude photos of herself. Her struggles with harassment continued for more than 6 years. (Curtis, 2018)

As we see, even in societies which are in terms of gender much more equal, for instance Australia (ranked 39 on the global scale in 2018 according to the World Economic Forum, 2018) women are finding it very hard to fight against pornographic deep fakes. For instance, laws against non-consensual usage of pictures or videos for pornographic purposes were only introduced into the Australian juridical system in 2018 after multiple victims have spoken out about their experience. Under these legislations convicted individuals could face jail time up to five years for distributing non-consensual intimate images and up to seven years of imprisonment if they are more than three times offenders. Also, according to civil penalties these crimes will be punishable to up to AU$105,000 for individuals and AU$525,000 for corporations. (Reichert, 2018) But, while these new laws might help persecute offenders in Australia, they will not be applicable to perpetrators who are overseas. It seems that once again nothing can stop perpetrators from countries without such copyright laws from acquiring someone's photos or videos online even if that person has a closed and private account. Adding to that the unwillingness of law enforcement and juridical institutions to help the victims, like in the case of Ayyub, then we have an environment where perpetrators will be able to operate freely and without consequence, helping the use of deep fakes to become commonplace.

When it comes to the victims it is quite alarming to see that most of these involuntary fake videos are of women known and unknown alike. This is specifically evident when we take into consideration the enormous amount of pornographic deep fakes of celebrities and unknown persons in recent years. While men are not an exception when it comes to deep fake exploitation, they mostly never end up in a fake porn video. Their faces are usually doctored into prank videos or videos which intend to sabotage their reputation in other ways. Although these all can be extremely harmful to a person, nevertheless it seems that the area of fake porn is reserved for women only. (Harwell, 2018) These videos have been "weaponized disproportionately against women, representing a new and degrading means of humiliation, harassment and abuse. The fakes are explicitly detailed, posted on popular porn sites and increasingly challenging to detect." (*Ibid*) These examples highlight the gendered dimension of the phenomenon: "[t]his has been the case for cyber stalking and non-consensual pornography, and likely will be the case for deep sex fakes." (Chesney and Citron, 2018: pp. 17-18) There is a growing number of websites and closed discussion forums which offer deep fakes for their users for the starting rate of $20 per fake. (Harwell, 2018) The targeted victims are mostly "women far from the public eye" referred to by anonymous users as "co-workers, classmates and friends." (*Ibid*)

Chesney and Citron (2018) stress that when victims find out that their faces have been superimposed into pornographic images and videos, the psychological pressure can be quite destructive. And it does not even matter what the creator intended to do with such a deep fake. Victims can feel an entire range of emotions from being scared, to being humiliated or even psychologically abused. Fake porn videos treat women like sexual objects forcing them into non-consensual virtual sex. (Chesney and Citron, 2018: p. 18) A more chilling effect of deep fake porn videos is that they can also "transform rape threats into a terrifying virtual reality. They send the message that victims can be sexually abused at whim." (*Ibid*) When talking about her first reaction to seeing the fake porn video about herself, Ayyub said it was altogether a devastating experience:

> "I just couldn't show my face. You can call yourself a journalist, you can call yourself a feminist but in that moment, I just couldn't see through the humiliation…I felt so embarrassed. The entire country was watching a porn video that claimed to be me and I just couldn't bring myself to do anything." (Ayyub, 2018)

On the other hand, Noelle Martin whose face was superimposed to dozens of porn videos and images over the years says that she felt like as if someone was stripping away her dignity and humanity. (Laschon, 2018)

Just how vulnerable women are when it comes to online sexual exploitation by deep fakes was demonstrated by yet another fake app called DeepNude released in June 2019. This latest app "undresses" women's pictures and by doing so it creates extremely realistic nude images of the victims. The targets are only women. The creator, who remains anonymous, claims he used more than 10 000 pictures of naked women to train the app's algorithms, because they were easier to find online. At this point the app will generate nude bodies of women even if you provide the algorithm a picture of a man. However, the creator does not rule out the possibility of creating an app that will undress men as well. Even though the DeepNude did not create real nude photos of actual women, the pictures can easily be mistaken for a real thing and cause irreversible damages to the victims. As we have seen already in the case of Martin and Ayyub, fake porn is a really powerful mechanism to intimidate, harass and silence women. (Hao, 2019) Due to the highly vocal online backlash the "DeepNude has now been taken offline, but it won't be the last time such technology is used to target vulnerable populations." (*Ibid*)

Online surveillance can have quite a deterring effect on victims. Knowing that their fake porn videos or images were shared and viewed by thousands of people will give additional reasons for the victims to retract, feeling defeated, humiliated and violated. This is especially true of vulnerable groups, who are online even more vulnerable. (Singh and Roseen, 2018) Although nowadays everyone can participate in the production of knowledge, inequalities still persist in society as well as within the digital environment. Thus, resulting in certain people having more power to share information about a certain matter than others do. This means that not all information is equally visible or accessible by everyone. (Hanell and

Salö, 2017) While being visible has some advantages. visibility is also considered to be "a double-edged sword: it can be empowering as well as disempowering." (Brighenti, 2007: p. 335) Vulnerable groups who are not in powerful positions in a society are also less able to participate in the production of knowledge. Thus, they are considered invisible, which means they are "being deprived of recognition." (*Ibid*: p. 329) However, the link between recognition and visibility is not always straightforward. According to Brighenti, thresholds of visibility play an important role in determining one's visibility and recognition in a society. This "fair visibility", as Brighenti calls it, can range from minimum to maximum, regardless of the nature of the criteria we use. Some people who find themselves below the lower threshold of fair visibility are considered "socially excluded". (*Ibid*: pp. 329-330) Like for instance, illegal immigrants or homeless people.

> "On the other hand, as you push yourself – or are pushed – over the upper threshold of fair visibility, you enter a zone of supra-visibility, or super-visibility, where everything you do becomes gigantic to the point that it paralyses you. It is a condition of paradoxical double bind that forbids you to do what you are simultaneously required to do by the whole ensemble of social constraints…Distortions *in* visibility lead to distortions in social representations, distortions *through* visibility." (Brighenti, 2007: p. 330, emphasis original)

Therefore, vulnerable individuals, like Ayyub and Martin, who are 'pushed' into the zone of super-visibility are unable to influence on their own terms their online narratives.

Another terrible outcome for the victims of such pornographic deep fakes is the huge amount of inappropriate and offensive sexual offers they get from strangers and sexual predators alike. Noelle Martin for instance noted that the first doctored sexually explicit pornography of hers which appeared on the internet was made using a picture she took when she was only 17. She says she was specifically targeted because people fetishized her.

(Laschon, 2018) Rana Ayyub had a similar experience. She recalls many people leaving her distasteful comments on Facebook talking about her “stunning body” while others didn’t even shy away of openly asking her how much money she charges for sex. (Ayyub, 2018)

Even if such deep fakes will be debunked it might be too little too late for the victims depicted in them. Apart from the inflicted psychological damage and online abuse, these deep fakes can cause individuals to lose their jobs or altogether to destroy any prospect for their current or future careers. The post-truth climate that we live in only increases the chances of such deep fakes to be distributed by thousands of people online regardless of their factuality, causing irreversible reputational damage to individuals in the process. (Chesney and Citron, 2018: p. 19) And once out there on the World Wide Web shared by thousands and thousands of people, these deep fakes will be hard if not impossible to remove. In 2018,

> “Google added ‘involuntary synthetic pornographic imagery’ to its ban list, allowing anyone to request the search engine [to] block results that falsely depict them as ‘nude or in a sexually explicit situation.’ But there’s no easy fix to their creation and spread.” (Harwell, 2018)

Martin knows exactly how this feels; even after years of fighting for justice she still encounters dozens of deep fakes with her image doctored into them just by googling her name. (Laschon, 2018)

To make the matter even worse, people’s confirmational biases and algorithmic processing raise the potential risks of the fakes being encountered by their future employers, colleagues, acquaintances and even romantic interests. (Chesney and Citron, 2018: p. 19) In the case of Ayyub, the pornographic deep fake about her was shared by many members of the Hindu nationalist political party, BJP, and was posted on their official fanpage, boosting its visibility so much so that in just mere few days the video was shared more than 40,000 times.

The whole country had access to her deep fake porn. (Ayyub, 2018) As it was already mentioned in the *Post-truth* section, the higher the visibility of the story, the more credible it is among people who encounter it. The fact that nowadays information is distributed in an algorithmic manner, and that people share it with like-minded people on social media platforms, has changed the way in which things become visible to us. (Hanell and Salö, 2017: pp. 154-156; Chesney and Citron, 2018: p. 13) Moreover, if certain misinformation or fake news targeting an individual becomes a viral phenomenon, it only gains more visibility and thus potentially causing even further harm. Yet it is not only high visibility that makes these videos believable. As mentioned, it is people's susceptibility to being easily swept away by rumors, lies and misinformation which reinforce their pre-existing beliefs and opinions that pose a real threat. (Davies, 2018: p. 7; Prego, 2017: pp. 20-21) Ayyub for instance, in one of her recent articles from 2019, talks about this kind of toxic environment present in India. She argues that the hatred and violence against ethnic minorities, especially Muslims, is spreading like a disease both online and offline as well. In addition, the ruling Bharatiya Janata Party and its president, Amit Shah, encourage this kind of behavior by publicly expressing their lack of respect and hatred towards Muslims every chance they get. In a tweet from April 2019, they stated that they would get rid of all the infiltrators of the country who are not Buddhists, Hindus or Sikhs. Later on, Shah took this promise a step further by vowing he will get rid of all the 'termites' by throwing them into the Bengali Bay. However, besides the growing nationalism and increased hostility towards Muslims and other minorities, what Ayyub finds even more alarming is that "many citizens have found this new language of hate liberating and acceptable." (Ayyub, 2019) And it is exactly this climate, which celebrates narratives that are in line with the ruling political ideology that has enabled Ayyub's deep fake porn to become so popular and credible among the public. (*Ibid*) Such a public that allows itself to “be blinded permanently” will threaten to destroy the country’s democratic system altogether. (*Ibid*)

4.2.Deep fakes and their potential to harm societies

Our technological capacities to create videos that can mimic reality quite accurately come at a time when society is already vulnerable to misinformation and fake news. The production of knowledge is no longer in the hands of (trusted) media companies, the academia and the government. Nowadays everybody can create their own versions of the truth and facts and distribute it to the rest of the world. Our confirmational biases play a role in the acceptance of misinformation, and algorithms help spread them further. (Chesney and Citron, 2018: p. 9) This also means that “[d]eep fakes are not just a threat to specific individuals or entities. They have the capacity to harm society in a variety of ways.” (Chesney and Citron, 2018: p. 20) The fact that they have not been utilized on a much greater scale yet does not mean that potential harm caused by them is an unrealistic scenario. Since they are indeed a very new phenomenon, we have yet to understand what specific risks they present to our society. However, given the enormous amount of fake news and misinformation campaigns in 2016, which contributed to the British people voting for Brexit and Trump being elected for president gives us some insight into what we might expect when it comes to deep fakes as well.

Chesney and Citron (2018) were among the first to try to outline the potential risks and dangers deep fakes pose to societies. They argue that there are a lot of beneficial outcomes as well from the creation of deep fakes, especially in arts, education and gaming. For instance, Hollywood already utilized deep fake technology to bring to life deceased actors for certain roles. The most famous examples of this are in the *Star Wars* franchise, where the characters played by the late Carrie Fisher and Peter Cushing were recreated with the help of this technology. Gaming industry is yet another area where deep fakes might come in handy. Gaming platforms, such as Nintendo Wii, already developed certain games where users can create their own customizable avatars. When it comes to education, Chesney and Citron argue that this technology could also be used to alter existing films, documentaries or shows for pedagogical purposes. (Chesney and Citron, 2018: pp. 14-16) For instance "[w]ith deep fakes, it will be possible to manufacture videos of historical figures speaking directly to students, giving an otherwise unappealing lecture a new lease on life." (*Ibid*: p. 14)

Despite all the beneficial possibilities of deep fake technology, it is their potential harm that we should be focusing on instead. This should be done considering our already vulnerable digital environment and current societal developments. Chesney and Citron (2018: pp. 16-21) point out that the threat presented by deep fakes to society has first of all systematic characteristics, since they have a potential to reach and harm all levels of society. For instance,

> "[t]he damage may extend to, among other things, distortion of democratic discourse on important policy questions; manipulation of elections; erosion of trust in significant public and private institutions; enhancement and exploitation of social divisions; harm to specific military or intelligence operations or capabilities; threats to the economy; and damage to international relations." (*Ibid*: p. 21)

However, there is no way of accurately predicting what a specific deep fake might do. The potential harm depends on the context of its creation and circulation, and it is only possible to understand its full impact once it had time to reach people. By that time of course some damages could be irreversible.

In May 2018 a deep fake of Donald Trump addressing the Belgian public emerged.[3] In this crudely doctored video the fake Trump is seen calling out on the Belgian public to

[3] The deep fake of Donald Trump addressing the Belgian public can be accessed through this link: https://hoax-alert.leadstories.com/3469396-fake-news-spa-on-twitter.html

urge their government to withdraw from the Paris climate agreement. The video was created by a Belgian left-wing political party called Socialistische Partij Anders (or sp.a) and posted on their official social media accounts. The video, with the intention to spark people's interest in issues of climate change, received hundreds of angry comments, many stating that Trump has no right at all in expressing his opinion about a Belgian political matter. (Schwartz, 2018)

One outraged tweet read: "Humpy Trump needs to look at his own country with his deranged child killers who just end up with the heaviest weapons in schools." (as cited by Schwartz, 2018) Another went even further throwing insults on the entire American nation: "Trump shouldn't blow so high from the tower because the Americans are themselves as dumb." (*Ibid*) Needless to say, the deep fake managed to provoke and anger parts of the Belgian public quite a lot. Even the fact that the deep fake was really badly doctored – the fake Trump's lips and the sounds were off sync and the whole video was of very low quality – did not help the situation at all. The people who watched it were genuinely convinced that Trump really did say those things and understandably they directed their anger towards him. But that was a mistake. (Schwartz, 2018)

If one watches the video very carefully, at the end the fake Trump himself admits that he is actually featuring in a fake video. However, at the exact moment when he says those words the sound level drops drastically and the Dutch subtitles which were present during the entire video disappear, making it very difficult for those Belgian people who do not understand English to realize that what they have seen is not actual footage of Trump. Also, those having been paying attention only to the subtitles may be forgiven for having missed this final part. As it was later revealed, the sp.a requested the hi-tech forgery from a production house which uses AI to generate highly realistic fake videos of people. However, when the video did not end up making the expected outcomes, the party was left to manage the situation. Their social media team was forced to explain over and over again to their outraged followers that it was just a silly prank video and nothing more. (Schwartz, 2018)

Although the intentions of the creators of the Trump deep fake were to enlighten the Belgian public about the immediate threat of global warming and to inspire them to act on it, the effects were something they did not anticipate. And that is the problem, because once out there on the World Wide Web deep fakes can have a life of their own. The members of sp.a defended their actions by stating that, given the low-tech quality of the video, they assumed that their followers will instantly know it is a fake and understand the hidden message in it, which will eventually lead them to sign the climate change petition that the fake video was meant to popularize. (Schwartz, 2018) This is where one of the true dangers of such deep fakes lies; the inability of creators to predict how people will react on them and what kind of consequences they might have on the wider society. What we are dealing with here is a clear discrepancy between the creators' intention with such a technology and its everyday usage and perception by the public. For instance, the Belgian deep fake stirred up some negative emotions among the followers of sp.a. Many thought that Trump was disrespecting their nation and mingling with their country's politics. By playing on people's emotions, a deep fake is able to motivate them to (re)act. Negative emotions such as fear and the feeling of insecurity are incredibly strong motivators. They can create instability in diverse societies, eroding the trust in civic and democratic institutions resulting in a state of nervousness. (Davies, 2018: p. 20) According to Davies, "[m]uch of this nervousness that influences democracy today is not simply because feelings have invaded a space previously occupied by reason, but because the likely sources and nature of violence have become harder to specify." (*Ibid*: p. 17) In this sense violence does not necessarily mean physical violence; it can also mean just a threat of violence, a sense of danger or the feeling of insecurity. Nonetheless, this nervous state can shape the public's opinion so much so that even if the deep fake is proved to be false by outsiders, the "belief in it becomes an article of faith, a litmus test of one's adherence to that community's idiosyncratic worldview." (Donath, 2016 as cited in Zuckerman, 2017)

People in general fall for fake news and share them much more often than accurate news. This is due to the fact that information travels faster on social media platforms "if it looks and feels true on a visual and emotional level." (Davies, 2018: p. 15) Another idea to consider is that humans tend to care more for negative and novel information, which can evoke stronger emotions like surprise and disgust. (Chesney and Citron, 2018: p. 12) "Negative information not only is tempting to share, but it is also relatively 'sticky.' As social science research shows, people tend to credit—and remember—negative information far more than positive information." (*Ibid*) Moreover, humans are predisposed to pay close attention to things that will stimulate them, for instance things that are violent, sexual, disgusting, embarrassing and even humiliating. (*Ibid*: p. 13) Therefore, it is no surprise that many Belgians who saw the video were attentive to the fake Trump's disrespectful behavior and mistook him for the real Trump, at the same time completely ignoring the fact that the video was of low quality or that the fake Trump himself stated at the end that the video is just a fake.

Reports about this Trump video made international headlines as well, but it did not cause any major reactions on the global scale and it was widely ignored by the Trump administration as well. However, this might simply be because by the time the news about the video reached the rest of the world it had already been debunked and news outlets were stating it clearly that it was a case of a fake video meant to be a practical joke.[4] Another explanation as to why this video did not have some bigger impact on the global scale is

[4] The Belgian deep fake featuring Trump was first debunked by the online news website called Lead Stories on 20th May 2018. (Schenk, 2018)

maybe the fact that the fake Trump was very similar in character to the real Trump. Especially when it comes to declaring climate change as nothing more but a hoax or calling on other nations to follow the American example in dealing with political issues. Many people are already used to this real Trump; therefore, encountering another video of him ranting about how climate change is only fake news will not be anything *new* to them. Another thing to consider is perhaps the content of the video: climate change. People are very motivated to avoid any immediate threats, like a barking dog, or a dangerous street. Yet when it comes to climate change, people are more difficult to be moved. This topic does not have a lot of emotional appeal for individuals who have on a daily basis their own personal problems to deal with and do not necessarily feel the direct impact of global warming. (Markman, 2018) In fact, “many effects of climate change are distant from most people.” (*Ibid*) Therefore, it is no surprise that the Belgian public was more focused on (the fake) Trump addressing their nation and commenting on their politics, rather than the issue of climate change.

But nonetheless, even if this was just a small-scale deep fake incident, the fact that it was created by a political party in order to persuade people to change their minds about a certain political issue makes this scenario quite alarming. Especially if we add to this “the nature of today’s communications environment” in combination with our confirmational biases and algorithmic networks – then we have a clear recipe for a potential disaster even on a more global scale. (Chesney and Citron, 2018: p. 19) The results of a fake video circulating can be more significant in cases where the stakes are higher, or the emotional appeal stronger, for example if instead of trying to inspire people to care about the environment, the intent of the video would be to make them believe there has been a terrorist attack. As Davies (2018: p. 123) points out, even just “[s]mall acts of transgression can have major political effects, if the right tool and target are carefully selected.”

The Belgian deep fake came out just a month after Buzzfeed published an article warning about the possible risks of using deep fakes in political campaigns. In order to demonstrate the capacity of such technology and for the sake of the argument, they created a deep fake, featuring fake Barack Obama trash talking about Trump.[5] For this they utilized a free application called FakeApp, which was released by the Reddit user, "Deepfakes", who also created the fake porn videos of celebrities in 2017 mentioned above. In their article, which also showcased the forgery, they argued that although the technology to create doctored videos is in its infant stage, requiring a fair amount of IT skills, time and fast computers, the potential for it to become more sophisticated and democratized is just around the corner. The article makes the claim that if soon anyone can make a fake video that defies reality, then there are pretty perilous times ahead of us. (Silverman, 2018)

At the same time there are still some who oppose these apocalyptic predictions, claiming that the technology is not so far advanced that deep fakes could actually pose a real threat to our society. An article which appeared in The Verge in March 2019 stated that "deepfake propaganda is not a real problem" and it won't become an issue in our nearest future. (Brandom, 2019) The author of the text argues that the predictions of deep fakes becoming a threat to our democratic systems have not materialized yet. He adds that the main reason for this is simple: it's just not worth the trouble. He claims that the algorithms that can detect the forgeries are widely available online; therefore, it can be easily proved that the doctored videos are fakes. One of his arguments is that doctored videos are not useful enough to the extent that troll campaigns are. (*Ibid*) "Most troll campaigns focused on affiliations rather than information, driving audiences into ever more factional camps. Video doesn't help with that; if anything, it hurts by grounding the conversation in disprovable facts." (*Ibid*)

[5] The deep fake created by Buzzfeed featuring Barack Obama talking badly about Donald Trump can be accessed through this link: https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed?utm_term=.bgWEL9pQjg#.fs3VQGmReq

According to him deep fakes are in general more dangerous for individuals, given the huge amount of fake porn of celebrities and unknown people alike circulating the internet. Nonetheless, he remains skeptical about them becoming a real threat to our society anytime soon. (*Ibid*)

However, one might disagree with this. If the public reactions to the badly doctored video of Nancy Pelosi and the Belgian deep fake are of any indication, we have something to worry about. The fact that there has not been yet any major societal disruption by a deep fake does not mean that it cannot happen tomorrow, a week from now, a month from now or in the future. It seems that "[e]ven at this early stage [of technological advancement of deep fakes] it's proving difficult for humans to consistently separate" them from reality. (Silverman, 2018) If a picture says more than a thousand words, then what about a video? "In some instances, the emotional punch of a fake video or audio might accomplish a degree of mobilization to-action that written words alone could not." (Chesney and Citron, 2018: p. 24) What we have seen so far are doctored images, altered videos and fake news intended to distort the reality and change people's opinion about certain things. We have yet to experience a deep fake with the capacity to disrupt our senses of reality that will have bigger social and political ramifications. This is a possibility, as even if our algorithms would be able to detect the fake, it will not be enough. (Vincent, 2019)

One piece of evidence for this is that already some badly doctored videos have managed to create quite a lot of friction on the American political scene. The best recent examples are the Nancy Pelosi case from 2019 mentioned above, and the doctored video of a CNN reporter, Jim Acosta, seemingly hitting a White House assistant in 2018.[6] The latter being even a more chilling episode, considering the fact that the White House Press Secretary

[6] The altered video of Jim Acosta can be accessed through the following link: https://www.forbes.com/sites/laurenaratani/2018/11/08/altered-video-of-cnn-reporter-jim-acosta-heralds-a-future-filled-with-deep-fakes/#d063663f6c1e

Sarah Sanders shared the forged video on Twitter, while defending the White House's decision to permanently revoke Acosta's press pass. The part of the original video where the CNN reporter is seen taking the microphone back from the White House assistant, who was determined to stop him from further questioning the president, was accelerated to appear that Acosta actually hit the assistant. The forgery was created by Infowars, an online conspiracy website best known for its 2012 Sandy Hook Elementary School shooting conspiracy, in which they claimed that the shooting that led to 28 people being killed, actually never happened. Keeping this in mind, the fact that the White House decided to back its decision of banning the CNN reporter by using a forged video created by a conspiracy website raises some serious concerns. At the same time, it also demonstrates the growing risks of deep fakes being weaponized in the name of politics. (Aratani, 2018)

What these examples also illustrate is just how much one society can become easily polarized by fake news, doctored videos and very recently deep fakes as well. These lies and misinformation that have crept into our information networks and democratic systems have managed to destabilize institutions and create fear and mutual suspicion among the population. (Davies, 2018: p. 22) Fake news and misinformation are spreading with enormous speed online, due to people's confirmational biases, algorithmic boosting and filter bubbles. These filter bubbles in particular contribute to the societal polarization even more. Social media platforms create an environment where individuals find themselves enclosed within an informative bubble that reinforces their personal beliefs and opinions while at the same time suffocating any other narratives that might challenge the status quo. (Prego, 2017: p. 20) As Chesney and Citron (2018: p. 13) mention, "[f]ilter bubbles can be powerful insulators against the influence of contrary information." Opposing narratives that enter these bubbles are immediately discredited, leaving the filter bubble intact in the process. (Prego, 2017: p. 20) "In this atomized world that is self-strengthening, it is actually a huge weakness because it is the perfect breeding ground for spreading fake news." (*Ibid*: p. 21) People will not consider fact checking the information they receive in their bubble since they genuinely believe it to be true, because it confirms their own beliefs about a certain topic. (*Ibid*) However,

> "[o]ne of the prerequisites for democratic discourse is a shared universe of facts and truths supported by empirical evidence. In the absence of an agreed upon reality, efforts to solve national and global problems will become enmeshed in needless first order questions like whether climate change is real. The large scale erosion of public faith in data and statistics has led us to a point where the simple introduction of empirical evidence can alienate those who have come to view statistics as elitist." (Chesney and Citron, 2018: p. 21)

One possible solution for this problem might lie in educating the public how to tell apart facts from fiction and increasing media literacy. According to danah boyd, in order to achieve this, we need to be very creative and develop a structural base for people to communicate with each other across divisions in a meaningful way. (boyd, 2017a; boyd, 2017b) That is,

> "[w]e need to enable people to hear different perspectives and make sense of a very complicated — and in many ways, overwhelming — information landscape. We cannot fall back on standard educational approaches because the societal context has shifted. We also cannot simply assume that information intermediaries can fix the problem for us, whether they be traditional news media or social media." (boyd, 2017a)

Apart from this there is also the legal side of deep fakes to consider. Lawmakers and experts already warn that deep fakes might hinder and disrupt the upcoming U.S. elections in the year 2020. During a House Intelligence Committee hearing in Washington on the 13th of June 2019, experts from various fields as well as politicians were discussing the potential dangers of deep fakes and ideas how to prevent them from creating widespread damages. (George, 2019) Using the recent doctored video of Nancy Pelosi as an example, they argued that the era of deep fakes will have "the capacity to disrupt entire campaigns, including that for the presidency." (*Ibid*) This might be possible given the fact that the public is already struggling to separate facts from fiction. One expert suggested that private tech companies need to step up their game and ban such content from their platforms. However, giving these companies' freedom to decide on their own what kind of content should be removed was seen as too risky a move. Danielle Citron, a University of Maryland Law professor who also attended the meeting, told the lawmakers that most of the legislation regulating the usage of online videos are decades old and need urgent revision. (*Ibid*)

However, even with new laws restricting the creation and distribution of such forgeries it will not be enough to combat the entire problem, not when those laws have no jurisdiction outside their country of origin. For instance, "U.S. officials determined Russia carried out a sweeping political disinformation campaign on U.S. social media to influence the 2016 election." (*Ibid*) It is not unimaginable that they or anyone else will not try in 2020 as well to meddle with the political campaigns of presidential candidates. For instance, a deep fake incriminating a political candidate favored by the public released a night before the elections could potentially flip the results in favor of the other candidate. Even if the fake video would be debunked and proven false, it might be too late. "When events are unfolding rapidly and emotions are riding high, there is a sudden absence of any authoritative perspective on reality." (Davies, 2018: p. xi) In the heat of the moment people might not stop to consider the facts before acting. (*Ibid*) Deep fakes can indeed function like a new way of falsely shouting fire in a crowded theater. All it takes is just a well-timed deep fake to tip the election in favor of one of the candidates, “particularly if the attacker is able to time the distribution such that there will be enough window for the fake to circulate but not enough window for the victim to debunk it effectively.” (Chesney and Citron, 2018: p. 22) This is also because “[m]ore than anything else, the dynamics that define the web — frictionless sharing and the monetization of attention — mean that deepfakes will always find an audience.” (Vincent, 2019) Moreover, depending on the context of the shared deep fake, the willingness of people to accept facts that are in line with their own opinions and beliefs will further credit video forgeries. After all, if there is already some doubt among the public, deep fakes will deepen the mistrust even more. This is what happens with believers in conspiracy theories: they are vulnerable to believe in many if they believe in one. (Barkun, 2016: p. 2) On the other hand, if the target audience is not chosen carefully, either the content of the video will not be very relevant for the population or the timing is not right, in which case the video can be forgotten within 24 hours. This means that apart from the content of the deep fake, its publication and circulation strategy is also key in determining its potential effects.

However, the effects of deep fakes do not necessarily have to happen overnight, as online attacks are often a form of “slow violence”. (Varis, 2018) “[A] violence that occurs gradually and out of sight, a violence of delayed destruction that is dispersed across time and space, an attritional violence that is typically not viewed as violence at all.” (Nixon, 2013 as cited in Varis, 2018) In other words, small triggers that occur here and there can gradually over time create an avalanche of mistrust and mutual suspiciousness among the public. These continuous online persuasions eventually wear people down and on a large scale help stir the public’s opinion on a certain political matter. (Varis, 2018) Modern political campaigners deploy this strategy very often. They are very aware that public opinion can be best swayed with small-scale interventions, which sometimes go unnoticed, rather than big formal public statements. (Davies, 2018: p. 13) Therefore, a carefully executed misinformation campaign over a certain period of time will eventually lead to a much-desired result. And once it manages to polarize a society, the threats to the democratic systems start to become visible. (Chesney and Citron, 2018: p. 29) “If the public loses faith in what they hear and see and truth becomes a matter of opinion, then power flows to those whose opinions are most prominent—empowering authorities along the way.” (*Ibid*) In such a vulnerable state when our society is suffering from truth and trust decay, authoritarian leaders will strive to exploit the public’s opinion even further. (*Ibid*) This is already happening all around Europe. What we are witnessing is the rise of populism, which has managed to divide a once more integrated union of European countries with open borders and a shared currency into smaller conflicted islets. Populist leaders, such as Viktor Orban in Hungary, use the vulnerability of the European Union and specifically the refugee crisis to widen the divide even more. His misinformation campaign specifically targets political opponents, the academia, the media, NGO’s, prominent individuals and basically everyone who does not support his populist ideas and who are in position to have an influence on society. (Shattuck, 2019) If the credibility of individuals and institutions who have the capacity to produce and verify knowledge and information are undermined, the public is left with no choice but to believe those whose who hold the power, resulting in the erosion of democratic systems. (Chesney and Citron, 2018: p. 29)

In this post-truth climate, deep fakes are exceptionally dangerous. To put it in perspective: this climate gave birth to various misinformation campaigns which contributed to Brexit and Trump becoming a president in the first place. And this was all possible because the public is already struggling to make a difference between facts and fiction. This environment is especially vulnerable to all kinds of fake news, conspiracy theories and misinformation, because people who encounter them will act first and start asking questions later or, in the worst case scenario, never at all. Yet the capacity of deep fakes to mimic reality and fool not just our senses but also our technological tools created to counter them go beyond any other forms of fake news, misinformation campaigns or lies. Even just one realistic video forgery could have mind-blowing effects for which our democratic systems are just simply not prepared for.

5. Conclusion

There always tends to be a general moral panic about new technologies, but very often it is a misplaced worry. (boyd, 2012) However, when it comes to deep fakes the public anxiety that follows is actually well-based. A technology that is capable to fool our senses and even software and algorithms created to counter it is something to be worried about. (Chesney and Citron, 2018: n.p.) And since we live in a post-truth era where people are unable to tell apart facts from fiction, and where emotions and personal beliefs often dictate our knowledge about the world, even a crudely doctored video could trick people into believing it is real. (Oxford Dictionaries, n.d.; McIntyre, 2018: chapter 1, paragraph 12; Llorente, 2017: p. 9)

Despite being a relatively new phenomenon, deep fakes already managed to gain quite a negative reputation in Western societies. (Chesney and Citron, 2018) This is due to the fact that so many known and unknown people have fallen victims to this technology. While there is no denying that deep fakes can have some beneficial uses like for instance in education, Hollywood or in the gaming industry as it was mentioned above, still the different possibilities of causing harm to individuals and societies alike is what we should be focusing on instead. After all, given our weak information networks and current societal developments which gave rise to misinformation campaigns and mutual mistrust among the public, our democratic systems are simply not prepared for deep fakes at all. (*Ibid*: 16) Yet surprisingly there have not been any major academic studies focusing on this particular phenomenon. Therefore, the aim of this thesis was to outline the possible social, cultural and political ramifications deep fakes might have both on societal and individual levels through a discussion of specific case studies. This final chapter presents the findings of the analyses based on these cases by summarizing the central points and providing additional suggestions regarding future directions.

When it comes to their potential to harm individuals, we need to talk about the different motivations behind deep fakes and their various outcomes for victims and their environments. The motivations behind each deep fake can be various, depending mainly on the creative imagination of its creator(s). However, based on the cases reviewed in this thesis we can conclude that thus far the majority of deep fakes has been created with the intention to shame and silence individuals, to take revenge on someone, for both personal and political reasons; and very often to uphold the values and norms in a society. This latter is specifically the case when it comes to shaming and silencing women in society. Although everyone can become a victim of such technological forgery, some are more vulnerable than others. (Singh and Roseen, 2018) The cases of many known and unknown women who fall victim to fake porn videos indicate that there is a gendered dimension of sexual exploitation by deep fakes which is very worrisome. (Chesney and Citron, 2018: p. 17) Women are among the most vulnerable groups, in certain societies also more than in others. This is also true for the online environment. Vulnerable groups who are already bullied offline have a higher chance of becoming victims of online abuses as well. Since the lines between our online and offline lives are blurred, due to the fact that digitalization has created an unpredictable and ever evolving technological landscape, online threats can easily be transferred to the offline world as well. (Singh and Roseen, 2018) And given the nature of our social media platforms, where everybody can upload and share all kinds of information very often without any or with few restrictions, this environment is rather ideal for organized and systematic bullying and abuse. (*Ibid*) This climate also allows perpetrators to be anonymous and stay hidden, making it even harder for the victim to succeed in seeking justice for the harm that has been done to her. Additionally, our legal systems which are lagging far behind our current technological development make this matter even worse, since the victims are left alone to deal with their own problems while the perpetrators are never even persecuted. (George, 2019; Curtis, 2018)

On the other hand, the outcomes of deep fakes can be also different for each individual. Deep fakes can cause victims to experience various psychological harms, such as fear, anxiety or when it comes to fake porn videos, victims can be also subjected to visualization of their own sexual harassment, which can be also described as non-consensual virtual sex. (Chesney and Citron, 2018: pp. 16-20) One of the common outcomes of deep fakes is the retreat of victims from social media platforms. (Singh and Roseen, 2018) This is specifically the case for pornographic deep fakes where the intentions are very often to shame and silence women and to preserve the status quo in a society. Like other forms of online abuse which target women, deep fake bullying can result in women refusing to speak out about their experience and in some cases they choose to remove themselves from the online environment altogether. Sadly, this is also not specific to women alone. Other vulnerable groups who might be targeted by deep fakes can have the same or similar experiences. (*Ibid*)

Although nowadays everyone can have access and share information online, due to the digital revolution that democratized the production of knowledge, not everybody has the same amount of power to talk about all kinds of topics. (Hanell and Salö, 2017) Some people who are marginalized offline can also be invisible online as well. However, being visible does not always go hand in hand with benefits. This is because visibility can be empowering as much as disempowering. The latter is especially true in cases where marginalized individuals are pushed over the fair visibility threshold into the zone of supra- or super-visibility. In this environment vulnerable individuals do not get to choose how they are represented online leading to distortions in their digital visibility. (Brighenti, 2007: pp. 330-335)

Reputational sabotage is another outcome that can be attributed to deep fake abuse. Visibility of all kinds of information has drastically changed due to digitalization that led to information online being distributed in an algorithmic manner and shared on social media platforms, often shaped by filter bubbles. (Hanell and Salö, 2017: pp. 154-156; Chesney and Citron, 2018: p. 13) Yet, this environment also increased the chances of certain misinformation and fake news to become a viral phenomenon, which could potentially cause harm to both individuals and societies. While some deep fakes might be short-lived and forgotten within 24 hours, others might have long-term effects on people. Especially reputational harm can be very dangerous since it can have major long-term ramifications in different domains of one's life. An individual who falls victim to a deep fake might lose his or her job, lose the support and trust of friends and family members, lose partners and spouses, or be cast out of the community altogether. In a worst case scenario victims might feel entirely humiliated and left all alone, blaming themselves for the situations there are in. This can make it even harder for them to recover and restore normalcy in their everyday life. (Chesney and Citron, 2018: pp. 18-20)

Moreover, it is not necessarily only the victim that will be affected by a deep fake. In some cases, deep fakes that target specific individuals can have wider implications to the public sphere as well, such as dividing society into fractions and even posing a threat to the health of our democratic systems. People who are close to the victim, like family and friends or their community, might suffer as well. In some instances, the aftermath of a deep fake can be also felt on a societal level. A deep fake intended to harm an individual can trigger frictions in society, further deepening possible existing societal division and causing potentially even major disturbances to its democratic systems. (Chesney and Citron, 2018)

Deep fakes' potential to harm societies have first of all systematic characteristics, due to the fact that their effects can be felt on all levels of society. Here as well we can talk about both various motivations on behalf of the creators as well as types of outcomes deep fakes might have. The motivations behind deep fakes can vary endlessly. A forgery might be created to shape public opinion about a certain political matter, to create instability in society, to deepen the mistrust of people in private and public institutions, or to disrupt entire societies. Some deep fakes are created with the intention to strengthen a public's belief in a conspiracy theory or to confirm their doubts about a certain political issue, while others are used to discredit prominent individuals or media outlets. And again, certain deep fakes can be simply created to draw the public's attention to certain problems which are of outmost importance to them. (Chesney and Citron, 2018: p. 21)

When it comes to perpetrators, they can remain anonymous, but they can also be announcing their involvement in the creation and/or distribution of deep fakes. This is very true in cases when the creators have something to gain by revealing their true identity. On the other hand, deep fakes might be created by individuals, but they can also be a part of state-organized campaigns.

The consequences of deep fakes on societal level also vary. Deep fakes among other things can lead to major changes in the course of elections, damages to international relations, being used as weapons for war or terrorism, threats to national security, disruptions of economic developments, loss of trust in public institutions, polarization of society and threat to the democratic system altogether. (Chesney and Citron, 2018: p. 21) While some effects of deep fakes might happen overnight, like for instance a deep fake intended to discredit a political figure posted a day before elections potentially impacting the way people vote, others might have long-term effects. These long-term effects can be described as "slow violence" to the public sphere or democracy, since they result in minor disruptions over a longer period of time, which is the case for instance with the polarization of society, eroding trust in institutions and disruptions to democratic systems. (Varis, 2018; Chesney and Citron, 2018: p. 29)

Deep fakes intended to disrupt the normalcy of democratic systems might cause a state of nervousness in a society. (Davies, 2018: p. xii) As digitalization resulted in the revolution of how information is produced and shared among people (Hanell and Salö, 2017: pp. 155-156), it also resulted in creation of a state of continuous and heightened alertness of individuals and societies alike. Information online spreads with enormous speed, especially misinformation and fake news. This is due to the fact that, in this climate people rely mostly on their feelings and personal beliefs disregarding the objective facts in the process. This public anxiety can result in people being scared, feeling insecure, all the way to angry outbursts of certain individuals or verbal aggression. This nervousness is what follows fake news and misinformation campaigns online. (Davies, 2018: pp. 7-17) In the worst case scenario people who were fooled by a fake video might be motivated to act out and take the matter into their own hands, which of course can lead to physical violence as well.

Polarization of society is often seen as a major blow to the health of democratic systems. Divisions that already exist in a society can be deepened by deep fakes even more. (Chesney and Citron, 2018: p. 29) Filter bubbles can further aggravate such divisions. By creating enclosed spaces where information is used to reinforce people’s own opinions and beliefs, they present a perfect breeding ground for spreading misinformation, fake news and deep fakes. Very often opposing information does not appear in such filter bubbles or even if it does, it is to discredit it, while at the same time strengthening the existing confirmational biases of people. (Chesney and Citron, 2018: p. 13; Prego, 2017: pp. 20-21) This is especially true due to the fact that humans are more interested in negative and novel information rather than real facts. (Chesney and Citron, 2018: p. 12) Moreover, information that feels and looks true will travel faster online, further contributing to the problem. (Davies, 2018: p. 15) And if members of a certain society fail to agree upon what is real and what is fake, then instead of finding solutions to problems the focus will be on deciding whether those problems are a real issue in the first place. A public that is unable to agree upon reality itself is very vulnerable to exploitations and manipulations. In this toxic environment, authoritarian figures holding the power will be able to dictate what is true and what not, resulting in the creation of authoritarianism that presents a risk to our democratic systems as we know them. (Chesney and Citron, 2018: pp. 21-29)

And while the outcomes of using deep fakes can also be very different, they can also be unpredictable and are not always something that the creators had intended to achieve. This is due to the fact that creators of deep fakes cannot always anticipate accurately how people might perceive and react on these forgeries, or how they circulate for instance online. In order for a certain deep fake to have the desired effect on society its content, the targeted audience and the timing needs to be carefully chosen and even then, there is no guarantee that things might turn out the way the creators wanted to. To be able to determine the potential implications of deep fakes we need to explore their content, their publication and circulation strategy as well as the reactions of people who encountered them.

When it comes to possible solutions that might stop deep fakes from causing harm, the fact that there are no laws yet in many countries around the world that allow to recognize deep fakes as a form of (serious) crime means that both individuals and our democratic systems are left defenseless against such attacks. However, even if more countries would introduce legislation tackling the misuse of private photos of people for the purposes of creating forged images and videos, the problem would not be completely solved. The reason for this is either that perpetrators can go unidentified or can operate from countries where laws against deep fakes do not exist yet. (Reichert, 2018; Curtis, 2018) On the other hand, once out on the internet, deep fakes would be hard if not impossible to remove, given the fact that many people can potentially view them and share them further within in a very short amount of time. The lack of supervision and gatekeepers on social media platforms also aggravates this issue. (Madeiros, 2017: p. 23) However, the question is, should social media platforms be the ones deciding what kind of content needs to be removed in the first place? Can we trust companies such as Facebook or Google to make objective and unbiased judgments about this matter? On the other hand, appointing third parties to supervise online content such as governments, experts in AI or international organizations might just lead to unwanted censorship or biased decisions, which again are dangerous outcomes. In the end the question still remains whose responsibility it is then to stop deep fakes from spreading online. A possible solution might include a joint action to tackle the problem. This means restrictions on social media platforms, legal frameworks to persecute perpetrators and most importantly increasing media literacy in connection with proper education of the public on making a clear distinction between facts and fiction, while at the same time focusing on building “social infrastructures” where people would be able to meaningfully engage with each other. (boyd, 2017a) These new infrastructures would be like bridges connecting diverse parts of society and at the same time reducing social polarization. (boyd, 2017b) And these last three solutions might just be the most important in countering the potential effects of deep fakes.

However, despite the above-mentioned conclusions and findings, certain limitations and flaws remain regarding this research. As it was already pointed out, it is impossible to make definite generalizations about the risks deep fakes pose to individuals and societies based on solely the case studies explored in this thesis. Although the cases give a detailed insight into specific aspects of deep fakes, they are still limited to their specific context. While in this thesis it was argued that we should be mainly focusing on the harmful effects of deep fakes, this does not mean that their beneficial uses should be ignored. On the contrary, future studies might find it interesting to explore these aspects of deep fakes as well. Given the many social, cultural and political ramifications of deep fakes, future studies in this field need to further explore the motivation behind such video forgeries and the potential outcomes in order to find ways to counter them or at least to minimize the damage they will create. Since this thesis mostly focused on women as the individual victims of deep fakes, future research papers could explore in more detail other vulnerable groups in connection to deep fakes. A possible comparison between the genders is also another interesting way to highlight even more the issue of the gendered dimension of this technology. Another interesting aspect to explore might be the consequences deep fakes can have on the victim's family members, friends, community, and on the public sphere. More case studies about the possible implications of deep fakes concerning society would also widen our understanding of this phenomenon. Also, an interesting topic to consider is the social, cultural and political ramifications of technological solutions, such as software or algorithms, designed to counter deep fakes. Finally, future studies could focus on finding solutions to tackle the looming challenges of deep fakes. This might include exploring current and proposed legislative measures, the responsibilities of social media platforms, governments and tech giants in deciding what kind of content should be removed from the internet, possibilities of educating the public about the harmful effect of deep fakes, increasing media literacy, and finding new ways to bridge the differences in a society by building a more connected and functional social fabric.

6. References

Aisch, G., Huang, J. and C. Kang. (2016). *Dissecting the #PizzaGate conspiracy theories*. Retrieved from https://www.nytimes.com/interactive/2016/12/10/business/media/pizzagate.html

Aratani, L. (2018). *Altered video of CNN reporter Jim Acosta heralds a future filled with 'deep fakes'*. Retrieved from https://www.forbes.com/sites/laurenaratani/2018/11/08/altered-video-of-cnn-reporter-jim-acosta-heralds-a-future-filled-with-deep-fakes/#56ebf9053f6c

Ayyub, R. (2018). *I was the victim of a deepfake porn plot intended to silence me*. Retrieved from https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAKGJ5jRAo6u7sRJLGcx3zlwLIo3vLyHj3PkFE_MotDg_Ls0n3XPxmauv7B9PP_vn8-U3mZn_recRZCommm0QA-2iJxpEFPJ9NLFxTYFwVMr81GqijH7mucQ2cFxplSBxkF2lyTbi5FeBotUaX8R2e9IDHAcPYY9SJ0-QuZtE9JIx

Ayyub, R. (2019). *You know India's democracy is broken when millions wait for election results in fear*. Retrieved from https://www.dailymaverick.co.za/article/2019-05-17-you-know-indias-democracy-is-broken-when-millions-wait-for-election-results-in-fear/

Barkun, M. (2016). Conspiracy theories as stigmatized knowledge. *Diogenes*: pp. 1-7.

Bell, E. (2019). *At White House press conferences, no questions allowed*. Retrieved from https://www.theatlantic.com/politics/archive/2019/01/donald-trump-continues-to-call-the-media-fake-news/579670/

boyd, m. d. (2008). *Taken out of context: American teen sociality in networked publics* [PDF version]. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1344756

boyd, d. (2012). *The power of fear in networked publics* [transcription of a talk]. SXSW. Austin, Texas Retrieved from http://www.danah.org/papers/talks/2012/SXSW2012.html

boyd, d. (2017a). *Did media literacy backfire?* Retrieved from https://points.datasociety.net/did-media-literacy-backfire-7418c084d88d

boyd, d. (2017b). *Why America is self-segregating*. Retrieved from https://points.datasociety.net/why-america-is-self-segregating-d881a39273ab

Brandom, R. (2019). *Deepfake propaganda is not a real problem*. Retrieved from https://www.theverge.com/2019/3/5/18251736/deepfake-propaganda-misinformation-troll-video-hoax

Brighenti, A. (2007). Visibility: A category for the Social Sciences. *Current Sociology*. 55 (3): pp. 323–342.

Carlsson, U. and R. Pöyhtäri (eds.). (2017). *The assault on journalism: Building knowledge to protect freedom of expression* [PDF version]. Retrieved from http://www.unesco.se/wp-content/uploads/2017/04/The-Assault-on-Journalism.pdf

Chesney, R. and D. K. Citron. (2018). *Deep Fakes: A looming challenge for privacy, democracy, and national security* [Draft version]. Retrieved from: https://ssrn.com/abstract=3213954

Collins, T. (2018). *Trump's itchy Twitter thumbs have redefined politics*. Retrieved from https://www.cnet.com/news/donald-trump-twitter-redefines-presidency-politics/

Curtis, C. (2018). *Deepfakes are being weaponized to silence women — but this woman is fighting back*. Retrieved from https://thenextweb.com/code-word/2018/10/05/deepfakes-are-being-weaponized-to-silence-women-but-this-woman-is-fighting-back/

Davies, W. (2018). *Nervous states: How feeling took over the world*. London: Jonathan Cape.

George, S. (2019). *'Deepfakes' called new election threat, with no easy fix*. Retrieved from https://www.apnews.com/4b8ec588bf5047a981bb6f7ac4acb5a7

Gillespie, T. (2014). The relevance of algorithms. In Gillespie, Tarleton, Pablo J. Boczkowski and Kirsten A. Foot (eds.) *Media technologies: Essays on communication, materiality and society. MIT Scholarship Online*: pp. 167-193.

Gooch, E. (2017). In pursuit of the truth. *UNO. The Post-truth Era: Reality vs. Perception*. 17: pp. 14-15.

Gordon, N. (2019). *Scarlett Johansson on how helpless she feels about her image being used in fake porn videos*. Retrieved from https://www.cosmopolitan.com/uk/reports/a25722524/scarlett-johansson-helpless-face-fake-porn-videos/

Hanell, L. and L. Salö (2017). Nine months of entextualizations. Discourse and knowledge in an online discussion forum thread for expecting parents. In Kerfoot, Caroline and

Kenneth Hyltenstam (eds.) *Entangled discourses. South-North orders of visibility*. London: Routledge: pp. 154-170.

Hao, K. (2019). *An AI app that "undressed" women shows how deepfakes harm the most vulnerable*. Retrieved from https://www.technologyreview.com/s/613898/an-ai-app-that-undressed-women-shows-how-deepfakes-harm-the-most-vulnerable/

Harwell, D. (2018). *Fake-porn videos are being weaponized to harass and humiliate women: 'Everybody is a potential target'*. Retrieved from https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/?utm_term=.5c9026480758

Harwell, D. (2019). *Faked Pelosi videos, slowed to make her appear drunk, spread across social media*. Retrieved from https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/?noredirect=on&utm_term=.e54e8a1b6041

Hergott, M. (2019). *A leap into the future: Generative adversarial networks*. Retrieved from https://medium.com/datadriveninvestor/a-leap-into-the-future-generative-adversarial-networks-96a780ed8ee6

Hollinger, J. (2017). *Trump, social media and the first Twitter-based presidency*. Retrieved from https://www.diggitmagazine.com/articles/Trump-Twitter-Based-Presidency

Lamb, K. (2018). *Indonesian government to hold weekly 'fake news' briefings*. Retrieved from https://www.theguardian.com/world/2018/sep/27/indonesian-government-to-hold-weekly-fake-news-briefings

Laschon, E. (2018). *Revenge porn crackdown announced by WA Government offers hope for victims*. Retrieved from https://www.abc.net.au/news/2018-06-28/revenge-porn-crackdown-announced-by-wa-government/9920560

Llorente. A. J. (2017). The post-truth era: Reality vs. perception. *UNO. The Post-truth Era: Reality vs. Perception*. 17: p. 9.

Lyons, K. and K. Lamb. (2018). *Sulawesi tsunami: Indonesia battles fake news as hoaxers spread panic*. Retrieved from https://www.theguardian.com/world/2018/oct/04/sulawesi-tsunami-indonesia-battles-fake-news-as-hoaxers-spread-panic

Madeiros, A. (2017). The danger of indifference to truth. *UNO. The Post-truth Era: Reality vs. Perception*. 17: pp. 23-25.

Markman, A. (2018). *Why people aren't motivated to address climate change*. Retrieved from https://hbr.org/2018/10/why-people-arent-motivated-to-address-climate-change

McIntyre, L. (2018). *Post-truth* [Kobo Aura version]. Retrieved from https://www.kobo.com/nl/en

Medrán, A. (2017). In the kingdom of post-truth, irrelevance is the punishment. *UNO. The Post-truth Era: Reality vs. Perception*. 17: pp. 33-35.

Mezzofiore, G. (2018). *No, Emma Gonzalez did not tear up a photo of the Constitution*. Retrieved from https://edition.cnn.com/2018/03/26/us/emma-gonzalez-photo-doctored-trnd/index.html

Mikkelson, D. (2018). *Was Emma González filmed ripping up the U.S. Constitution?* Retrieved from https://www.snopes.com/fact-check/emma-gonzalez-ripping-up-constitution/

Mondal, S. (2017). *Why was Gauri Lankesh killed?* Retrieved from https://www.nytimes.com/2017/09/13/opinion/gauri-lankesh-india-dead.html

O'Grady, S. (2018). *An Indian journalist has been trolled for years. Now U.N. experts say her life could be at risk*. Retrieved from https://www.washingtonpost.com/news/worldviews/wp/2018/05/26/an-indian-journalist-has-been-trolled-for-years-now-u-n-experts-say-her-life-could-be-at-risk/?utm_term=.fddc4b5773e0

Oxford Dictionaries, (n.d.). *Post-truth*. Retrieved from https://en.oxforddictionaries.com/definition/post-truth

Prego, V. (2017). Informative bubbles. *UNO. The Post-truth Era: Reality vs. Perception*. 17: pp. 20-21.

Reichert, C. (2018). *Australia passes 'revenge porn' legislation*. Retrieved from https://www.zdnet.com/article/australia-passes-revenge-porn-legislation/

Schenk, M. (2018). *Fake news: Belgian social democrat party uses faked Trump video in climate change campaign*. Retrieved from https://hoax-alert.leadstories.com/3469396-fake-news-spa-on-twitter.html

Schwartz, O. (2018). *You thought fake news was bad? Deep fakes are where truth goes to die*. Retrieved from https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth

Shattuck, J. (2019). *How Viktor Orban degraded Hungary's weak democracy*. Retrieved from https://theconversation.com/how-viktor-orban-degraded-hungarys-weak-democracy-109046

Silverman, C. (2018). *How to spot a deepfake like the Barack Obama–Jordan Peele video*. Retrieved from https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed?utm_term=.bgWEL9pQjg#.fs3VQGmReq

Singh, S and D. Roseen. (2018). *Online, vulnerable groups only become more vulnerable*. Retrieved from https://www.newamerica.org/weekly/edition-229/online-vulnerable-groups-only-become-more-vulnerable/

Snapes, L. (2018). *Bataclan survivor Jesse Hughes calls March for Our Lives 'pathetic'*. Retrieved from https://www.theguardian.com/music/2018/mar/26/eagles-of-death-metal-jesse-hughes-march-for-our-lives-bataclan?utm_term=Autofeed&CMP=twt_b-gdnnews#link_time=1522064213

Solsman, E. J. (2019a). *Deepfakes may ruin the world. And they can come for you, too*. Retrieved from https://www.cnet.com/news/deepfakes-may-try-to-ruin-the-world-but-they-can-come-for-you-too/

Solsman, E. J. (2019b). *Samsung deepfake AI could fabricate a video of you from a single profile pic*. Retrieved from https://www.cnet.com/news/samsung-ai-deepfake-can-fabricate-a-video-of-you-from-a-single-photo-mona-lisa-cheapfake-dumbfake/

Sundar, S. (2016). *Why do we fall for fake news?* Retrieved from https://theconversation.com/why-do-we-fall-for-fake-news-69829

Swaine, J. (2017). *Donald Trump's team defends 'alternative facts' after widespread protests*. Retrieved from https://www.theguardian.com/us-news/2017/jan/22/donald-trump-kellyanne-conway-inauguration-alternative-facts

The Star Online. (2018). *Police urge public not to fall for fake news*. Retrieved from https://www.thestar.com.my/metro/metro-news/2019/03/22/police-urge-public-not-to-fall-for-fake-news/

Varis, P. (2018). *What is the wholesome internet? How wholesome memes became a trend*. Retrieved from https://www.diggitmagazine.com/column/wholesome-internet-memes

Vincent, J. (2019). *Deepfake detection algorithms will never be enough*. Retrieved from https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work

Wagner, J. (2018). *Trump renews attacks on media as 'the true Enemy of the People'*. Retrieved from https://www.washingtonpost.com/politics/trump-renews-attacks-on-media-as-the-true-enemy-of-the-people/2018/10/29/9ebc62ee-db60-11e8-85df-7a6b4d25cfbb_story.html?utm_term=.77979ed9510a

World Economic Forum, (2018). *The global gender gap report 2018* [PDF version]. Retrieved from http://www3.weforum.org/docs/WEF_GGGR_2018.pdf

Zainal, Z. (2007). *Case study as a research method*. Retrieved from https://www.researchgate.net/publication/41822817_Case_study_as_a_research_method

Zarzalejos, A. J. (2017). Communication, Journalism and Fact-checking. *UNO. The Post-truth Era: Reality vs. Perception*. 17: pp. 11-13.

Zuckerman, E. (2017). *Fake news is a red herring*. Retrieved from https://www.dw.com/en/fake-news-is-a-red-herring/a-37269377