Deepfakes and extreme beliefs

An ethical assessment

Jack Esselink

Table of contents

ABOUT THE AUTHOR **4**

PREFACE **5**

ABSTRACT AND KEYWORDS **6**

LIST OF ABBREVIATIONS **7**

1. INTRODUCTION **8**

2. METHODOLOGY **11**

2.1 Case study method 11
2.2 Thought experiment method 13
2.3 My position as researcher/ thinker 14

3. DEEPFAKES DEFINED **16**

3.1 Deepfakes, how it works 16
3.2 Baudrillard's simulacra-model 21
3.3 Deepfake applications 22
3.4 Epistemic consequences of deepfakes 24
3.5 Deepfakes, conspiracy theories and extreme beliefs 27
3.6 Will empathy save the *Homo Syntheticus*? 31

4. INTRODUCTION TO REFLECTIVE EQUILIBRIUM **33**

4.1 Reflective Equilibrium in detail 33
4.2 Criticisms 38
4.3 Reflective Equilibrium in the technological context of this thesis 40
4.4 Conclusion 41

5. CASE STUDIES **42**

5.1 Justification for case study selection 42
5.2 Case 1. Alleged Deepfakes 43
5.3 Case 2. Deepfakes today 45
5.4 Case 3. Deepfakes ten years from now 47
5.5 Reflective Equilibrium method applied 50

6. CASE STUDY 1. FOOLED BY FAKES **52**

6.1 Ethical issue 52
6.2 Stakeholders and interests 52
6.3 Considered judgments 53
6.4 Moral principles and background theories 55
6.5 Reflection 58
6.6 Conclusion and lessons learned 60

7. CASE STUDY 2. SEEING IS (NOT) BELIEVING **62**

7.1 Ethical issue 62
7.2 Stakeholders and interests 62
7.3 Considered judgments 64
7.4 Moral principles and background theories 68
7.5 Reflection 72
7.6 Conclusion and lessons learned 76

8. CASE STUDY 3. WILL FAKE BE THE NEW REAL? 78

8.1 Ethical issue 78
8.2 Stakeholders and interests 78
8.3 Considered judgments 80
8.4 Moral principles and background theories 81
8.5 Reflection 85
8.6 Conclusion and lessons learned 87

CONCLUSION 89

BIBLIOGRAPHY 94

APPENDICES 114

Appendix A. Animated photo created by *DeepNostalgia*. 115
Appendix B. Picture of havoc wreaked by rainfall in Germany (July 2021). 116
Appendix C. *Migrant Mother*, picture by Dorothea Lange (1936). 117

About the author

Jack Esselink has worked in the software industry for more than two decades and has held various international commercial and technical positions at large software companies like IBM. Currently, he is a speaker and trainer in the field of Artificial Intelligence (AI) and ethics at his own company *Studio Pulpit* (www.studiopulpit.com). This master thesis has been written as a part of Jack's MA in Theology and combines his interests for ethics, extreme beliefs, and technology. If you are interested in learning more about these topics or have any questions about this master thesis, than you can contact Jack via his personal website www.jackesselink.nl.

Preface

I have always enjoyed exploring the impact technology has on our society. There is a reciprocity in the interaction humans have with technology, or as the adage goes "we shape our tools and thereafter they shape us" (Culkin 1967: 70). We live in interesting times, especially since the COVID-19 pandemic, where the use of digital technology has accelerated, and it amplifies human intentions. If you couple this with the emergence of smart technologies like Artificial Intelligence (AI) and virtual reality, it is expected to have a profound impact on our society and will shape our (near) future. The advent of AI powered tools is not unproblematic and exposes important ethical questions we need to answer as a society. For this thesis I have focused on the recent phenomenon of deepfakes in the context of extreme beliefs, which to me is the embodiment of the 'ideal' combination where technology, ethics, epistemology, and theology meet.

For this research I have met and talked to many interesting researchers, philosophers, authors, technologists, historians, and ethicists. They all sharpened my vision on what deepfakes are and what their potential impact is going to be in the context of extreme beliefs, philosophy, and ethics. I would like to thank dr. Rik Peels for his thesis supervision, and I would also like to thank all the members of the *Extreme Beliefs* research group for their interesting perspectives, helpful suggestions, and great articles they shared on our monthly research group meetings. I would also like to thank dr. Rob Compaijen for helping me to better understand the concept of reflective equilibrium and dr. Quassim Cassam for pointing me in the right direction regarding deepfakes, ethics, and philosophy. Finally, I would like to thank all the interviewees for taking the time and effort to help me out on this fascinating journey. This thesis is the culmination of my quest into ethics, (extreme) beliefs and technology and I have really enjoyed it and I hope to continue this in my work as a consultant in Ai and ethics for both (Christian) charity and other organizations. I hope the reader will enjoy this thesis as much as I have enjoyed writing it.

Abstract and keywords

Deepfakes are a nascent technological phenomenon that is expected to have a profound impact on our society. In this thesis I will conduct an ethical assessment using the reflective equilibrium method for the use of deepfakes in the context of groups holding extreme beliefs. This research will *not* provide a normative ethical evaluation but will expose what moral principles are at stake. It is based on three case studies that are situated in Belgium and the Netherlands in the years 2020, 2021 and 2031. The selected case studies each represent an increasing distance towards reality which is modeled after the different stages in Jean Baudrillard's simulacra-model (1994). The exposed moral principles will vary from *pro tanto* personal principles, like freedom of speech and informed consent, to *pro tanto* societal principles, like climate justice, transparency, epistemic authority, credulity, and the *pro tanto* obligation to do no harm. It is argued that for future case studies that involve an ethical assessment of deepfakes, these moral principles are useful to properly contextualize deepfakes as a social phenomenon. Deepfakes should not be considered as a new and isolated technical category, but as a technology wrapped in a broader, social context in our society that will amplify existing sociological trends like the diminished trust in epistemic authorities. Deepfakes can and will be weaponized by groups holding extreme beliefs and should be seen as the latest technology manifestation in the creation of disinformation and propaganda. In general mis- and disinformation will lead to an epistemic deterioration of our information environments (De Ridder 2021) and deepfakes will only accelerate and amplify this. The insights from this research will help both researchers in academia and the general public to take a broader, more nuanced, and contextualized view to assess the moral impact of deepfakes and it will help to inform the public debate around deepfakes and increase media literacy.

Keywords: deepfakes, synthetic media, ethics, reflective equilibrium, extreme beliefs, conspiracy theories, artificial intelligence, case studies, Jean Baudrillard, simulacrum.

List of abbreviations

AI	Artificial Intelligence
EU	European Union
NVR	No Vaccine Repeat[1]

[1] Name of fictitious group holding extreme views. This is used in the thought experiment in chapter 8.

1. Introduction

"I am not Morgan Freeman" is the first sentence Morgan Freeman speaks in a video of which you might think this *is* the real Morgan Freeman, however once the video progresses the virtual Morgan Freeman explains that he is not even a human being but a synthetically created virtual person who wants to welcome the viewer "to the era of synthetic reality" (Diep Nep 2021). This video is a good example of a video that has been edited in such a professional manner that the end-result looks very real but is completely fake. In popular media as well as academic research these kinds of videos are called *deepfakes*[2] and I find them a fascinating, emerging technological phenomenon that is expected to have far-reaching philosophical, legal, and moral implications for our society in terms of how we perceive truth and reality. Deepfakes are a very new phenomenon; the word *deepfake* only exists since 2017 (Meckel & Steinacker 2021; Giansiracusa 2021: 46) and scholarship around this topic is gradually picking up from various disciplines like computer science (e.g., Westerlund 2019; Lanham 2021; Langguth et al. 2021), media studies (e.g., Meckel & Steinacker 2021; Dobber et al. 2021; Vaccari & Chadwick 2020), political science (e.g., Schick 2020; Barari et al. 2021), law (e.g., Chesney & Citron 2019; Langa 2021), theology (e.g., Anderson 2019, 2021) and philosophy (e.g., Floridi 2018; Rini 2020; Harris 2021).

I am very interested in the impact new technologies, like deepfakes, have on the beliefs humans hold and what this means for morality in a society. For this thesis I have conducted research in the context of the multi-year research program *Extreme Beliefs* which runs at the Vrije Universiteit (VU) Amsterdam, and which is supervised by dr. Rik Peels. The goal of this research program is geared towards developing "a new normative-theoretical framework for better understanding and explaining fundamentalism" (Extreme Beliefs 2021). My research for this thesis focuses on the ethical implications that deepfake technology may have on extreme beliefs in a society.[3] This has led to the following research question:

What moral principles are at stake in the use of deepfakes in the context of groups or people holding extreme beliefs in the Benelux[4] in the years 2020, 2021 and ten years in the future?

What makes this research unique is the contextual approach using real life case studies in the context of extreme beliefs. Most of the published ethical and

[2] A proper definition of deepfakes will be provided in chapter 3 *Deepfakes defined*.
[3] A proper definition of extreme beliefs and fundamentalism will be provided in chapter 3 *Deepfakes defined*.
[4] The Benelux is an acronym for three neighboring countries in Europe: Belgium, the Netherlands and Luxembourg.

philosophical academic articles approach deepfake technology as a generic phenomenon and describe the potential epistemic and ethical impact it may have in the future. Much of this research is based on hypothetical presumptions of potential epistemic consequences of using deepfake technology. I have not come across any academic research that does an ethical assessment of the use of deepfakes in real-life political cases or in the context of extreme beliefs. This can be explained by the fact that there are not many documented examples to date about the use of political deepfakes (Ajder et al. 2019).[5]

For my research I have selected three case studies that describe a real-life or potential real-life situation where deepfake technology is being used or is allegedly being used. The structure of this thesis is as follows: the first three chapters are supporting chapters that will set the stage for the moral evaluation of the three selected case studies. In the first chapter I will discuss the used methodology (case studies and thought experiments) and my position as a researcher/ thinker in the moral reflection. The next chapter outlines the concept of deepfakes, its epistemic consequences and examples of applications. In addition, the simulacra-model (Baudrillard 1994) that is underpinning the selected case studies will be introduced and an introduction on extreme beliefs and conspiracy theories in the context of deepfakes will be provided. The third supporting chapter will give an extensive introduction into the method of *reflective equilibrium*, which is the ethical method used to conduct the moral evaluation for each case study. In chapter five each case study will be introduced followed by a selection justification and how the reflective equilibrium method will be applied for each case study. The subsequent three chapters are an account of the applied reflective equilibrium process for each individual case study in which the underlying moral principles, my perspective as a thinker[6] and their mutual coherence are described. Each chapter will end with a short conclusion and lessons learned section. In the final *Conclusion* chapter, the thesis will be summarized, and the research question will be answered accompanied by other findings, lessons learned and recommendations that the research revealed.

Deepfakes are expected to inundate the internet and it will become more and more difficult to distinguish between fake and real content. This will have huge impact on how people will perceive reality, and this can e.g., fuel conspiracy theories and that's why research on deepfakes will be relevant for people both inside and outside

[5] Ajder et al. (2019) found that 96% of all documented deepfakes in 2019 are related to non-consensual pornography.

[6] The role of a thinker is crucial in conducting ethical assessments based on reflective equilibrium. See §2.3 for more background on my position as a thinker and chapter 4 for more background on the role of a thinker in the reflective equilibrium process.

academia. It is relevant for academia, especially researchers in humanities and researchers studying extremism and fundamentalism, as it provides an ethical analysis of real-life cases where deepfakes are used and they can apply the lessons learned and recommendations for future cases they may have at hand. Outside academia, Deepfakes are expected to have great impact on society in both good and evil ways. On the good side, e.g., creative organizations can use deepfakes[7] to create more tailor-made content for less money by recording one training video and distribute this in twenty different languages using the same actor.[8] The evil side of using deepfakes is covered extensively in this thesis. At the time of writing of this thesis,[9] deepfake videos are now often used as a meme,[10] but it is expected that it will contribute to an *infocalypse*[11] in the world. The philosopher Jessica van der Schalk even calls this the biggest danger to society of our time.[12] Research on this topic is therefore also important outside academia. Research on deepfakes will help organizations and policy makers help better understand and assess this phenomenon. The insights and recommendations from this research are primarily targeted towards the context of extreme beliefs but will also be helpful beyond this.

[7] I am using the term *deepfake* here for both benign and malicious applications of the technology. I will argue in chapter 3 of this thesis that the term *deepfake* is only associated with malicious applications and that benign applications are denoted by the term *synthetic media*.

[8] An example of this is the British former football player David Beckham who speaks nine different languages in this anti-malaria promotion video (Malaria Must Die 2019). In this video the movement of his lips are synced with the language he is speaking using deepfake technology.

[9] This thesis is written over the summer and fall of 2021.

[10] Memes are ideas or 'culture carriers' packaged as short movies or animated pictures (GIF) that are distributed via social media (Van Doorn et al. 2021: 154).

[11] The term *infocalypse* is defined by Schick (2020: 7) as "the increasingly dangerous and untrustworthy information ecosystem within which most humans now live."

[12] Statement made on the *Future Affairs* podcast of the Dutch newspaper *NRC* (Felix Meritis 2020).

2. Methodology

In this chapter I will provide an explanation and justification for the different philosophical methods to answer the research question of this thesis. I have used three methods: case studies, thought experiments and reflective equilibrium of which the latter is the methodological focal point for this thesis. I have created two case studies and one thought experiment that provide the input for using the ethical reflective equilibrium method. Since the latter is such a crucial component in this thesis, it deserves its own chapter (chapter 4). I have selected the reflective equilibrium method for two reasons: (i) this method is used in the broader context of the Extreme Beliefs research program for ethical analysis and (ii) it is one of the most widely used methods in (applied) ethics. In the first and second section of this chapter the case study and thought experiment method will be explained and justified. The final section contains an account for my position as a researcher.

2.1 Case study method

The case study method is one of the most widely used methods in social sciences because of its natural appeal to readers of the conducted research (Stake 2009) and its ability to make abstract theoretical concepts easier to grasp. Thomas & Myers (2015: 7) provide a good working definition for a case study:

> Case studies are analyses of persons, events, decisions, periods, projects, policies, institutions or other systems which are studied holistically by one or more methods. The case that is the subject of the inquiry will be an instance of a class of phenomena that provides an analytical frame – an object – within which the study is conducted and which the case illuminates and explicates.

The key components for a case study following this definition are the *subject* and the *object*. It is easy to assume that the subject should be representative for a larger population of similar cases as usually happens when selecting a sample from a larger population when conducting quantitative research. According to Thomas & Myers this does not apply to case studies since "the *subject* will be selected because it is an interesting or unusual or revealing example through which the lineaments of the *object* can be refracted" (2015: 56), in other words, the selected subject provides a lens to the researched object. The object acts as the analytical backdrop that provides the concepts, context, and purpose of the research.

The 'free format' that is associated with case studies can invoke uncertainty among research who think case studies aren't methodologically sound and because of its

open-endedness that 'anything goes' (see Thomas & Myers 2015 for a discussion). Thomas & Myers (2015: 66) propose a typology for designing case studies that will help pre-empt this objection. The typology combines the subject/ object components with the purpose, analytical approach and the various processes and is schematized below in figure 1.

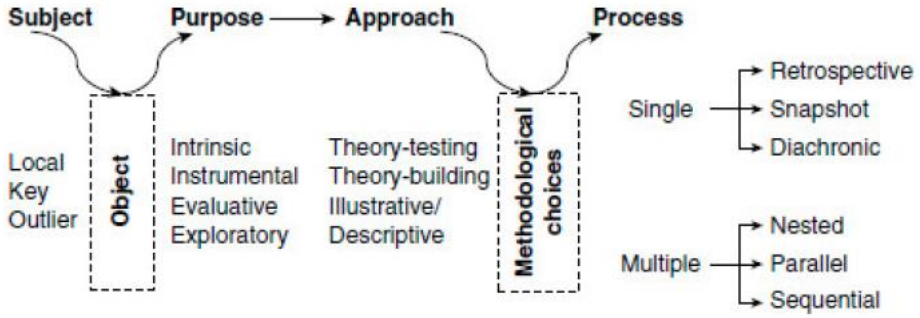

Figure 1 A typology of case study (source: Thomas & Myers 2015: 64)

In this thesis I will use three case studies of which two are based on events that have taken place recently. The third case study will take place in the future and therefore have used a thought experiment (see next paragraph) but that will have similar characteristics as the other case studies. Based on Thomas & Myer's typology the cases used in this thesis will have the following characteristics: these are *key subject* cases with an *exploratory* purpose using an *illustrative/ descriptive* approach describing a *single* phenomenon based on a *snapshot* time-use. Snapshot time-use explores cases in a "defined period of time" and multiple cases can be juxtaposed based on this time dimension (Thomas & Myers 2015: 63).[13] The time-dimension is important to the selected cases in this thesis since the progression in time clearly demonstrates the increasing impact deepfake technology has had and is going to have.

The use of case studies has received its share of criticism from the academic community of which lack of generalization is the most common one. This is based on the misconception that all academic research is based on scientific induction and deduction and tries to develop a general description of all applicable cases, however, the intention of qualitative research is generally not to generalize (Rubiano A. 2021). According to Stake (2009: 19) case studies are "epistemologically in harmony with the reader's experience and thus to that person a natural basis for generalization." The reader will be able to recognize the patterns and outliers presented in the cases and will be able to apply these lessons learned in different contexts. Stake (2019: 22) calls this epistemological process *naturalistic generalization* where the reader builds up

[13] For a detailed exposition of the *Retrospective* and *Diachronic* time-use categories see Thomas & Myer (2015: 63).

knowledge through experience. I agree with Stake that well selected cases have properties that can be recognized and applied to other cases. In the selection of the cases for this thesis I have tried to make them recognizable for the readers of this thesis. After reading this thesis they will be able to recognize similar cases and be aware of the ethical implications that are associated with it and apply the lessons learned from these cases in their own context.

2.2 Thought experiment method

Thought experiments can be thought of as imaginary or fictitious cases and are frequently used in philosophy (e.g., Dever 2016; Brown & Fehige 2019; Weisberg 2016) and ethics (e.g., Dancy 2021; Brun 2017; Walsh 2011). In this thesis the focus will be on conducting an ethical thought experiment which Walsh (2011: 469) defines as "to consider what would be the case *morally* if the particular state of affairs described in the imaginary scenario were actual. In effect we are asked to determine the moral status of that hypothetical state." Many ethical thought experiments are modally formulated (Dancy 2021; Brun 2017) using possible worlds and *contingent* moral properties of these possible worlds. Let *T* be an ethical thought experiment that describes an imaginary state of affairs *S* that takes place in a possible world *W'* where *M* is the moral status of *S* in *W'*. The following are the three most common applications of *T* used in (applied) ethics (Walsh 2011): (1) to provide a counterexample in which *S* contrasts a moral case in the actual world *W*, (2) drawing attention to *M* in *S* because these features are morally salient and, (3) help us to provide a new perspective on controversial, stale moral cases like e.g., abortion. Thought experiment *T* in the third case study intends to paint a picture what the impact of deepfake technology would be ten years from now, in other words, it provides a description of *S* and *M* in *W'* where *W'* is ten years ahead of *W*. The description of *S* and *M* in *W'* will be an extrapolation of *S* and *M* in *W*.[14] The input for this comes from interviewing experts on deepfakes, reading books and articles on this topic and the experience of myself being an expert in the field of *Artificial Intelligence* (AI)[15] for the last twenty years. In Brun's functional taxonomy of thought experiments (2017: 198-202) *T* would be a *heuristic*, *illustrative* thought experiment that is used for describing and explaining a potential state of affairs *S* and its moral status and properties *M* of which the result will be investigated independently using the reflective equilibrium method.

In addition to providing the content for the third case study thought experiments will also be used when conducting the reflective equilibrium process in all the case

[14] *W* being the actual world at the time of writing this thesis (fall 2021).
[15] See §3.1 for a proper working definition of AI.

studies. During the reflective equilibrium process constructive and destructive *epistemic* thought experiments (Brun 2017: 199) will be used to provide additional considered judgments to the reflective equilibrium process.[16] These thought experiments enrich the reflective equilibrium process, but the thinker needs to be aware that considered judgments that are based on thought experiments cannot be the sole justification for an equilibrium.

The use of thought experiments in ethics has also received its share of criticism (see e.g., Walsh 2011 and Brun 2017 for a discussion). It can be argued that the description of *S* and *M* in *T* is being insufficient since not enough context is provided of *S* in *W'* for a sound moral judgment. Another objection found is that thought experiments could lead to unreliable results (Brun 2017: 206) because the description of *S* and *M* will never take place in *W.* A related objection to the former is that in some thought experiments the distance between *W'* and *W* is simply too large that this leads to "morally outlandish stories" that should not be justified to use in ethics (Walsh 2011: 468). I think thought experiments are a useful tool in the ethicist's toolbox, but the ethicist needs to be aware that the purpose of the thought experiment is and that it meets what Walsh (2011: 479) calls the *contingency constraint*, which is that the thought experiment is keeping it in context with the moral case that is being investigated so it does not change the topic. In sum, thought experiments are a legitimate and justified tool to use in ethics and will provide input for the third case study and support the reflective equilibrium process in this thesis.

2.3 My position as researcher/ thinker

In qualitative research the role of the researcher or thinker is not objective (Rubiano A. 2021) since she brings her own biases, preferences and experiences into the research which can influence the research outcome.[17] In this section I would like to account for my position as researcher by providing a description of my background, worldview and my reflection on the thinker-role for this thesis with a description of the potential conflicts.

Personal worldview

My professional background is in AI and I have worked for twenty-five years as a consultant in the IT industry. My personal worldview is that I am an active Pentecostal-charismatic Christian who believes that the role of my belief is something that is the core of my personal identity and informs every decision I take in my life. One of the

[16] The reflective equilibrium process and the notion of considered judgments will be explained in chapter 4 *Introduction to Reflective Equilibrium*.
[17] In this thesis the terms 'researcher' and 'thinker' will be used interchangeably.

major reasons why I am pursuing a master's in Theology, is that I would like to combine my professional background with my personal worldview. As technology plays a more and more important role in our society, I see it as my personal conviction that I would like to help organizations, both Christian and non-Christian, in navigating this technological world in an ethical fashion.

Relationship to research

As mentioned above, I am very interested in the intersection of faith, ethics and technology which has informed the choice of my research topic. I have deliberately chosen to take an epistemological and ethical approach in my research, and not a theological approach for a couple of reasons. The first reason is that the added value of the outcomes of my research is higher, as it can be applied by anyone dealing with a ethical case in the area of the use of deepfakes. Second reason is that I would like to apply the obtained knowledge in my work in advising companies, churches, and Christian organizations on technology matters. The third, and last, reason is that the epistemic-ethical approach is the core method of the Extreme Beliefs project of which my research is a subproject.

Potential conflicts

The first potential conflict would be the ignorance of my personal ethical view and try to work under the assumption that I could conduct an objective ethical evaluation. The reflective equilibrium method challenges the researcher to bring her own ethical beliefs into the research as well as other perspectives. By using the reflective equilibrium method I am pre-empting both this potential conflict and the inverse, which is letting my personal ethical view prevail over other ethical stances. In addition to this I will apply the notions of being *embedded and embodied* as can be found in the work of Paul Ricoeur (Moyaert 2014: 33). These notions helped me to understand my own position when I compiled the case studies for this thesis. This will also be helpful to pre-empt another potential conflict which is the conflict that I think these people are 'wappies' which could make me feel morally superior to them.[18]

[18] The term 'wappies' is a popular term used in Dutch media to describe people who believe in certain conspiracy theories. The term is a morally laden term that entails othering and implies a moral superiority of the person using this term.

3. Deepfakes defined

In this chapter the concept of deepfakes will be defined, explored, and explained. Nina Schick (2020:6) has defined deepfakes as a malign manifestation of synthetic media, which are "media (including images, audio and video) that is either manipulated or *wholly generated* by AI." In the first section this definition will be unpacked and explained in the context of data and algorithms. In the second section Jean Baudrillard's (1994) simulacra-model will be explained as the theoretical backdrop for this chapter and the selected case studies. In §3.3 several applications of deepfakes will be covered and in the subsequent section (§3.4), the epistemic consequences of using deepfakes will be explored. The penultimate section outlines the relationship between deepfakes, conspiracy theories and extreme beliefs and in the final concluding section the concept of empathy will be explored as a mechanism for humans to thrive in a deepfake era.

3.1 Deepfakes, how it works

Most scholars who investigate the topic of synthetic media and deepfakes tend to use the words 'deepfake' and 'synthetic media' interchangeably (e.g., Paris & Donovan 2019; Floridi 2018; Johnson & Diakopoulos 2021; Meckel & Steinacker 2021; De Ruiter 2021). They follow regular and social media which tend to prefer the word deepfake over synthetic media. I agree with the distinction that Schick (2020: 7) makes, where deepfakes are "any synthetic media that is used for mis- and disinformation purposes" to denote the malign application of deepfakes.[19] Schick's definition offers a clear view on the various (technical) components that are involved in creating, editing, and disseminating deepfakes. An additional definition of deepfakes is provided by De Ruiter (2021: 2) and shows why deepfakes can be very problematic: "Deepfake technology refers to machine learning techniques that can be used to produce realistic looking and sounding video or audio files of individuals doing or saying things they did not necessarily do or say." I think both Schick's and De Ruiter's definitions are too narrow and should be extended with synthetically generated text (e.g., Dale 2021; McGuffy & Newhouse 2020; Giansiracusa 2021). In sum, a deepfake can be any type of media that is synthetically generated or manipulated using AI technology and its goal is to create fear, uncertainty, and doubt (Fallis 2020).

[19] I prefer to use the term 'synthetic media' over 'deepfakes' because of the negative connotation the latter has in the public opinion. However, since the focus of this thesis will be predominantly on the negative impact of these technologies, I will use the word deepfake in this thesis.

AI, data and algorithms

Using technology to manipulate media and therefore altering reality has been very common throughout history (for examples, Kessler & Schäfer 2018; Schick 2020; Langguth et al. 2021; Paris & Donovan 2019). What sets deepfakes apart from traditional manipulation techniques is the use of AI to generate content that is (almost) indistinguishable from reality.[20] The two terms that make up the acronym AI, 'artificial' and 'intelligence,' are philosophical ambiguous terms, so it goes without saying that many different definitions of AI can be found (see e.g., Hauser 2021; Bringsjord & Govindarajulu 2018; Russel & Norvig 2021). A good working definition of AI is provided by Gordon & Nyholm (2021): "AI is the use of machines to do things that would normally require human intelligence." In other words, deepfakes are media that are synthetically generated by an artificial agent, being software most of the time, of which the generated output looks like it has been created by a human. A good example of this is the painting *Portrait of Edmond de Belamy* (see figure 2) which was created using AI (Stephensen 2019) by the French art collective *Obvious*.[21] This painting was sold at a Christie's auction in New York for $432,000 (Jones 2018).

Figure 2 Painting 'Portrait of Edmond de Belamy' created using AI (source: Stephensen 2019: 21).

The painting has been generated by a specific type of AI algorithm that is called a *Generative Adversarial Network* (GAN) which was invented in 2014 by Ian Goodfellow

[20] De Ruiter (2021: 2) uses the term 'machine learning' in her deepfake definition. For this thesis my presupposition is that machine learning and AI are the same, and both terms can be used interchangeably. For this thesis I will use the term 'AI' by default. However, for AI researchers both terms can mean different things but have a huge overlap. It goes beyond the scope of this thesis to explore these differences (see e.g., Bringsjord & Govindarajulu (2018) for a discussion).

[21] The art collective *Obvious* explains in a blogpost how they use AI to generate art (Obvious 2018).

(Goodfellow et al. 2014). In order to use this algorithm, it must be trained on a large dataset that contains examples of the medium that one wants to create.[22] For example, if one wants to create a picture of a dog, then the GAN needs to be trained using a large collection of dog pictures. With the rise of using social media during the last fifteen years and other online services on the internet, it has become very easy to create or collect a dataset that can be used to train a GAN algorithm. Combined with the advent of cloud computing[23] during the last decade, the entry barrier to use GAN technology has dramatically lowered and has led an explosion of synthetically generated media. As one can imagine, the use of GAN technology has led to a huge increase of deepfakes (Schick 2020: 30-31; Meckel & Steinacker 2021: 14) and the synthetic results are completely indistinguishable from reality (Fletcher 2018) as the example in figure 3 shows; this is a GAN generated picture of a person that does not exist.

Figure 3 Example of a fake person generated by a GAN (source: www.thispersondoesnotexist.com). Picture generated on 5 October 2021.

Despite it has become much easier to create deepfakes, it still takes quite a lot of effort and skills to create a deepfake video that resembles reality. A good example of this are the deepfake Tom Cruise-videos (see figure 4) that are created by the Belgian visual effects-artist Chris Ume which he posts on the social media platform *TikTok* on a regular basis. Ume stated in an interview (CognitionX 2021) that for the creation of these videos he uses an actor, a Tom Cruise lookalike, who plays the videos and that

[22] For an in-depth technical overview of the use of GAN's, see Lanham (2021).

[23] Cloud computing can be defined as "By using virtualized computing and storage resources and modern Web technologies, cloud computing provides scalable, network-centric, abstracted IT infrastructures, platforms, and applications as on-demand services. These services are billed on a usage basis" (Baun et al. 2011). Providers of cloud computing like Google, Amazon Web Services and Microsoft offer cheap ways to rent computer infrastructure for which you pay what you use and without having to invest computer infrastructure upfront. The computing workload runs on the provider's infrastructure that sits in large data centers that are spread across the globe and users use an internet connection to access and use this.

he needs a lot of computing power to swap the actor's face with Tom Cruise's face. On top of that he is able to create this level of quality videos because of his background knowledge and experience as a visual effects-artist in the Hollywood film industry. In short, to create high quality deepfake videos (still) requires special IT-skills, lots of computing power and a large training dataset.

Figure 4 Screenshot of a deepfake Tom Cruise video on TikTok by Chris Ume (source: Ume 2021).

Spectrum of deepfakes and cheapfakes

The vast majority of deepfakes that are created nowadays are of a (much) lesser quality than the deepfake Tom Cruise videos. These videos are generated using smartphone apps like *ReFace*, *Avatarify*, *Face Swap Live* or *Wombo*, and become also more and more available in social media apps like *SnapChat* or *TikTok*. The main application of these apps is a *face swap*, which replaces someone's face with a different face. Also, more and more apps will bring a still image to life by creating a movie of it.[24] In addition to deepfakes, which are created using technologically advanced AI software, other forms of media manipulation are being used that require simple software (not powered by AI) or no software at all. This type of media manipulation which already existed long before deepfakes (Harris 2021) is coined as *cheapfakes* (Paris and Donovan 2019: 2) or *shallowfakes* (Langguth et al. 2021: 4; Giansiracusa 2021: 49).[25] A good cheapfake example is a video in which the American politician Nancy Pelosi is slurring and appears to be drunk (Diakopoulos & Johnson

[24] This became very popular when the online genealogy platform *MyHeritage* launched the *DeepNostalgia* feature that brought old photos to life by making the person on the photo smile or wink (Mahan 2021). See appendix A for an example.

[25] For this thesis I will use the term 'cheapfakes.'

2020: 2; Westerlund 2019: 43; Greengard 2020: 18). The video went viral in 2019 but it turned out that the creator decreased the speed of sound of the video to make Pelosi look bad (Paris & Donovan 2019: 30).

Paris & Donovan (2019: 10) have published an overview of the cheapfake-deepfake spectrum (see figure 5) of technological sophistication. The spectrum moves from easy to create, and almost no expertise required cheapfakes on the right, to sophisticated, AI based deepfakes on the left.

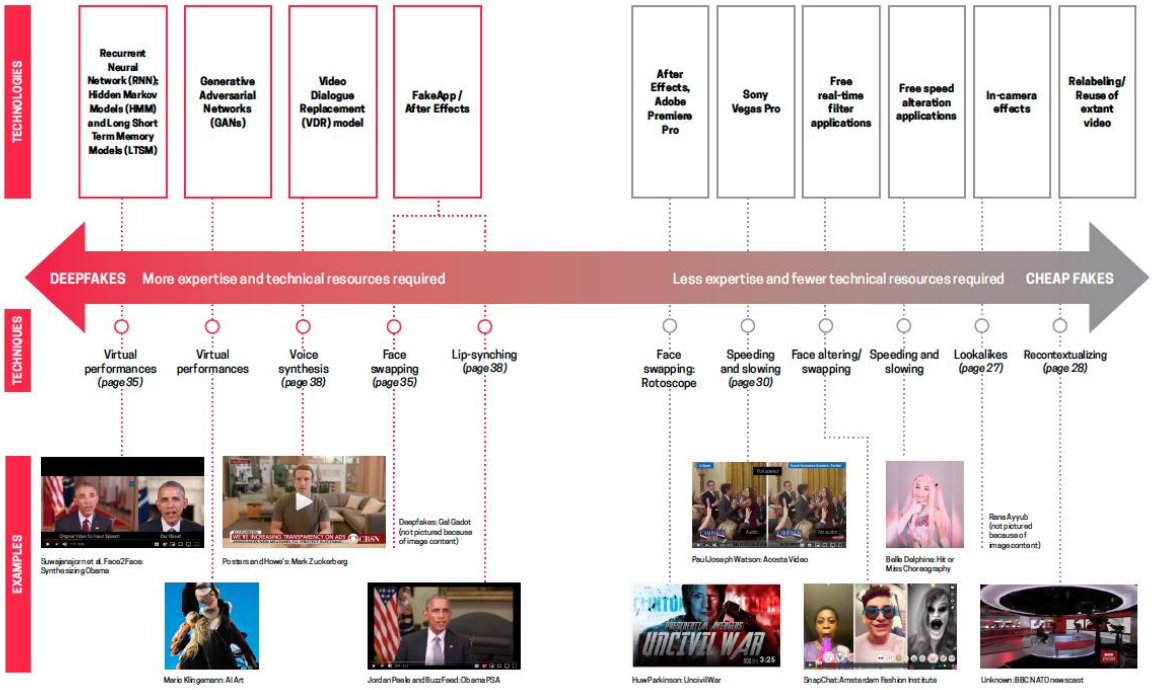

Figure 5 The deepfake-cheapfake spectrum including examples (source: Paris & Donovan 2019: 10).

It is expected that deepfake technology will become easier to use and more technologically advanced in the near future, in other words, it will become much easier to create deepfakes that are currently on the left end of the cheapfake-deepfake spectrum (Westerlund 2019). The consequence of this democratization of deepfakes (Anderson 2019: 8), or what Fletcher (2018: 456) calls the "game-changing factor," is that the scope, scale, and sophistication of deepfakes will increase exponentially. To illustrate this growth: people working in the synthetic media industry expect that in 2030, 90% of all online video content will be synthetically created (Schick 2020: 34).[26] In short, deepfake technology will become easier and available to more people in the (near) future. This will have enormous consequences for our society, and it is expected that fake and real will intertwine.

[26] This number comprises both malign deepfakes and benign synthetic media applications.

3.2 Baudrillard's simulacra-model

The notion of mixing fake and reality has already been modeled more than forty years ago by the French philosopher Jean Baudrillard in his simulacra-model. This model will be useful in this thesis as a theoretical framework that provides (1) a fake-reality taxonomy that underpins the selected case studies in this thesis and, (2) explains the use of deepfakes in conspiracy theories and extreme beliefs (see §3.5). In his book *Simulacra and Simulation* Baudrillard (1994)[27] describes how in our post-war Western society, reality is moving from a world where reality is being constructed by signs and symbols that *represent* this reality, to a world where reality is constructed by signs and symbols that are *completely disconnected* from reality. These symbols and signs in the latter stage are called *simulacra* (Baudrillard 1994: 6) and could be seen as copies without an origin or what the Italian philosopher of technology Floridi (2018) would call *ectypes*. In the end, simulacra become so ubiquitous that they become more important than reality itself and these simulacra have become *hyperrealities* (Harambam 2020: 145) and reality becomes a *simulation* (Baudrillard 1994: 1).[28] Baudrillard describes four successive phases how signs and symbols, which Baudrillard refers to as the *image*, develop towards a simulacrum. Baudrillard (1994: 6) writes "the successive phases of the image:

> It is the reflection of a profound reality;
> It masks and denatures a profound reality;
> It masks the *absence* of a profound reality;
> It has no relation to any reality whatsoever: it is its own pure simulacrum."

The four phases help to describe to what extent the artificially generated artefact,[29] what Baudrillard calls the image, is a faithful copy (De Jonge 2021: 13). The first phase is a true representation of reality, and an example could be a photo in the newspaper that shows the havoc that was caused by floods in Germany in July 2021.[30] The second phase is a perversion of reality which Baudrillard (1994: 6) calls "the order of maleficence" where the relationship between the image and reality has been obfuscated and put out of context. An example of the latter is footage showing Muslim immigrants attacking a Catholic church in France during mass (see figure 6) that was posted on a Facebook page in France in 2018, which was viewed 1.2 million times the after it was posted. However, the content in the video was taken out of context; fact checkers found out that the immigrants were not attacking the church at all but were

[27] Baudrillard's book was first published in 1981 in French as *Simulacres et simulation*.
[28] It is not a coincidence that a copy of the book *Simulacra and Simulation* is visible at the start of the film *The Matrix* (Harambam 2020: 213).
[29] In the context of this thesis this can be artificially generated pictures, photos, video, audio or text.
[30] See appendix B for an example.

protesting against a proposed bill that would make it harder to apply for asylum in France (Van der Linden & Roozenbeek 2020: 148).[31]

Figure 6 Screenshot of video posted in the Facebook group 'News World' (source: Van der Linden & Roozenbeek 2020: 148).

The third phase "masks the *absence* of profound reality" (Baudrillard 1994: 6) and denotes that the subject represented in the image has no relationship with reality, but humans may still be able to interpret this as being reality. This is the phase of deepfakes like the aforementioned example of Tom Cruise (figure 4). The fourth phase is pure simulacrum when there is "no relation to any reality whatsoever" (Baudrillard 1994: 6). Good examples of this are the aforementioned de Belamy-painting (figure 2) and the fake person in figure 3. In sum, according to Baudrillard (1994: 6) the world is moving from a world of "theology of truth" to the hyperreal world of "simulacra and simulation."

3.3 Deepfake applications

Since the concept of deepfakes is very new[32] and deepfake technology is developing and expanding at a rapid pace, there is not yet a comprehensive academic

[31] Despite the negative words used by Baudrillard to describe the second phase, there are not only negative examples. De Jonge (2021: 13-14) refers to an example of an iconic picture, called *Migrant Mother* (see appendix C) that was taken during the Great Depression in the United States. This picture became a representation of the hard times during the Great Depression rather than a reference to the woman in the picture.

[32] The word 'deepfake' is first mentioned on 2 November 2017 on the internet platform *Reddit* where someone started a discussion forum 'r/deepfakes' (Schick 2020: 25). In academic circles the concept of deepfakes was discussed in 2016 at a conference where researcher Justus Thies and his colleagues presented their research on real-time face capture and re-enactment (De Ruiter 2021: 4).

taxonomy of deepfake applications. The most common application reported for deepfakes is non-consensual porn (Schick 2020; Giansiracusa 2021; Rini 2020). Ajder et al. (2019) report that 96% of all deepfakes found were based on face-swap non-consensual porn of which most were the faces of female celebrities. The amount of deepfakes has been doubling approximately every six months and in an analysis conducted in 2020, by Ajder and his colleagues (mentioned in Giansiracusa 2021: 49), they found that the vast majority of people targeted in deepfakes (88.9%) came from the entertainment industry (including 21.7% from fashion and 4.4% from sports). Only 4.1% of the targeted people came from the business world and 4% from politics. These figures indicate that the main application, beyond non-consensual porn, is for humor and entertainment purposes. Deepfakes are used to create humorous or satirical memes that are shared on social media. Deepfake technology also leads to malpractice where it is harmful. Law professors Chesney and Citron (2019) have come up with a list of potential deepfake harms which are distributed in two categories: harms to organizations/ individuals and harms to society. The elements of the first category are blackmail, exploitation and sabotage. Chesney and Citron (2019: 1776) come up with several potential examples of the latter, like a fake audio clip might reveal criminal activities of a politician on the evening before the election takes place. Criminals already have used deepfake audio to mimic the voice of CEO of a German company who phoned the CEO of a subsidiary company to make a fraudulent money transfer (Brewster 2021; Westerlund 2019: 43; Stupp 2019). The potential harms to society that are described by Chesney and Citron (2019) range from distortion of democratic discourse, eroding trust in institutions to undermining diplomacy, national security and journalism. A potential example they (2019: 1776) provide is a fake video of an Israeli official saying something inflammatory that is starts riots in neighboring countries and has drastic political and diplomatic consequences. There are not many examples to date where deepfakes are used in a political setting, but a few have been documented, like the 2020 satirical address delivered by a deepfake Queen Elizabeth on Channel 4 (Giansiracusa 2021: 52), former US president Obama, impersonated by actor Jordan Peele, who mocked President Trump (Westerlund 2019: 43; De Ruiter 2021: 6). The last example is a video of the Gabonese president Ali Bongo who gave his annual New Year's address to his people after having not being present in the media for a couple of months due to a stroke. In the video it is clear that Bongo is still suffering from the stroke, but his opponents stated directly that the video was a deepfake. This caused an unstable situation in Gabon and a week after the president's address the military attempted a coup which ultimately failed (Rini 2020: 6; Giansiracusa 2021: 55-56; Breland 2019).

Deepfakes have been discussed extensively, both inside and outside, of academic literature. Despite that our society has not been inundated with deepfakes to date

which has not led to an infocalypse yet, governments and institutions are beginning to create deepfake policies. A good indication that deepfakes are taken seriously is an official alert that has been published by the FBI in March 2021 in which they state that "Malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations in the next 12-18 months" (FBI 2021). Maras and Alexandrou (2019: 257) expect that in the future deepfakes will be used more and more for e.g., revenge porn, bullying, fake evidence in courts, terrorist propaganda, fake news and market manipulation. In closing, despite the focus of this thesis is on deepfakes, I would also like to mention the various benign applications that synthetic media technology brings to bear (for examples see e.g., De Ruiter 2021; Giansiracusa 2021; Westerlund 2019; Kwok & Koh 2021; Kerner & Risse 2021).

3.4 Epistemic consequences of deepfakes

The use of deepfakes is expected to have enormous consequences for our society from an epistemological and psychological perspective. Philosophers like Regina Rini (Rini 2020; Rini & Cohen 2021), Don Fallis (2020), Catherine Kerner and Mathias Risse (Kerner & Risse 2021) and Adrienne de Ruiter (2021) have started to write about the impact deepfakes are expected to have from a philosophical and ethical point of view. In this section I would like to focus on the following epistemic consequences of deepfakes that in my opinion are most relevant for the context of this thesis: hermeneutics of suspicion, acquisition of false beliefs, reality apathy and the liar's dividend.

Hermeneutics of suspicion

By default, humans tend to trust what they read, see, and hear and what other people tell them, in real life or mediated by technology. This trust-default (Hancock & Bailenson 2021: 150) is one of the epistemic cornerstones of our society and is a necessary and sufficient condition for human communication and collaboration. In our current society, video is becoming the most important form of communication (Schick 2020: 34): we consume and produce videos and, more and more, we use videos on platforms like Youtube and Facebook as our main source of news updates.[33] Humans are wired with a *realism heuristic* (Meckel & Steinacker 2021: 18) that enables them to naturally believe video and audio that look or sound right. This cognitive bias, referred to by psychologists as *processing fluency* (Schick 2020: 21), helps us to trust what we hear and see by default, so we don't lose our ability to act, which might happen in case humans doubt everything (Anderson 2019). Deepfakes may cause to

[33] It is estimated that in 2022, 82% of Internet traffic will consist of streaming video and video downloads. Roughly 70% of the world's population, 5.6 billion people, is expected to have a smartphone and mobile internet connectivity in 2023 (Schick 2020: 21).

swap this trust-default from a *hermeneutics of faith* to what Anderson (2019: 2) calls a new *hermeneutics of suspicion*,[34] which makes people more suspicious to accept videos they watch as true. If it is true that deepfakes lead to a suspicious mindset towards the veracity of media, then it follows that this will have psychological and epistemic consequences like denial, fear, apathy, uncertainty, or skepticism. There are already examples where videos were misclassified as deepfakes that turned out to be real non-faked footage.[35] Vaccari and Chadwick (2020) found that people will not be completely fooled by deepfakes but being exposed to deepfakes has increased their uncertainty about media in general. Anderson (2019: 16) argues that a new hermeneutics of suspicion may be a good skill to have in an era that is dominated by deepfakes as this "present us an opportunity to reexamine our broader engagement with humans (and computers) online" and this will help "to develop digital systems that promote truth, empathy, and genuine depth."

Acquisition of false beliefs

According to Don Fallis (2020) the main epistemic threat that deepfakes pose is acquisition of false beliefs by people who perceive the false content of deepfakes to be true. Deepfakes do not only generate and justify false beliefs, but especially deepfakes may prevent people from acquiring true beliefs (Fallis 2020: 3). The epistemic cost that deepfakes incur is expected to increase when the amount of deepfakes will increase. There are already examples in real life that exemplify this epistemic consequence of deepfakes, each varying in epistemic cost. The first example is a video of the Canadian singer Justin Bieber who truly believed the deepfake video of Tom Cruise playing guitar (see figure 4). Bieber complimented the real Tom Cruise with his guitar skills on social media until someone notified him that he complimented the deepfake Tom Cruise (Thalen 2021). The second example took place in April 2018 in India where a deepfake video of two men on a motorcycle kidnapping a child, went viral on the social media platform WhatsApp and caused nationwide panic which resulted in many weeks of mob violence that killed at least nine innocent people (Vaccari & Chadwick 2020: 1). These examples demonstrate that deepfakes will carry less information about the topic they depict (as per Baudrillard's simulacra-model) and that the amount of deepfakes will increase the cost of acquiring true beliefs (Fallis

[34] Both terms have originally been coined by the French philosopher Paul Ricoeur (Anderson 2019; Moyaert 2014; Jasper 2004). According to Ricoeur a hermeneutics of suspicion acts as the counterpart of a hermeneutics of faith. Where a hermeneutics of faith tries to interpret the true meaning of media. A hermeneutics of suspicion, by default, questions its meaning and looks "beneath the surface for repressed or suppressed significance" (Anderson 2019: 2).

[35] A good example would be the first case study in this thesis (see §5.2 and chapter 6) which discusses the case where a fake spokesperson of the Russian political activist Navalny spoke with members of Dutch Parliament. During the conversation the fake-spokesperson started to show weird behavior and the MP's initially thought they were dealing with a deepfake, however, later it turned out they were pranked by two Russian comedians (e.g., Verhagen 2021; Roth 2021).

2020). The acquisition of beliefs in our society is often technologically mediated by an apparatus (Kessler & Schäfer 2018) which generates additional trust in the objectivity of the state of affairs it represents, compared to testimonies that represent a state of affairs. In our society the medium video has the highest epistemic trust status. According to Rini (2020: 10) video recordings generate *perceptual* knowledge which has a "stronger presumptive authority" than *testimonial* knowledge. When a person *P* is delivering a testimony *T*, the recording of *T* brings about a testimonial practice that will make *P* to speak with sincerity and competence; this is the *epistemic backstop* (RIni 2020) function of video. Because of advances in media technology our society has benefited from the epistemic backstop function of recordings, but deepfakes may diminish this. This may lead that our society goes back to old testimonial practices like eyewitnesses or newspapers (Rini 2020) and testimonial injustice[36] (Fricker 2007) of marginalized groups might increase.

Reality apathy

Reality apathy, also referred to as infocalypse, is a state of affairs *S* where person *P* gets exposed to such a large amount of deepfakes that it simply takes too much effort for *P* to determine whether the content of video *V* is real or not, and *P* simply sticks to her prior beliefs *B* (Toews 2020; Westerlund 2019; Schick 2020). If *S* is true for all persons in a society, then this society is unable to operate on a trust-default and all the basic ground rules that underpin a working society would need to be renegotiated in a renewed social contract. Schick (2020: 9) argues that *S* continually evolves and would lead to a society where it becomes more and more difficult to have a common shared reality, or consensus, on how to perceive and interpret the world. In this situation various psychological and technological effects will influence *P* in how she will maintain and reinforce *B*. The *illusory truth effect* (Meckel & Steinacker 2021: 15) states that repeated exposure to *V* will reinforce how *P* will maintain or revise *B* based on the content of *V*, even if *V* turns out to be wrong. This is often combined with a *confirmation bias* (Kahneman 2011; Chesney & Citron 2019) which states that *P* will seek evidence and confirmation based on *B* and this will make her maintain *B*. The algorithms of internet platforms, which are optimized to maximize engagement, will recommend, and expose content to *P* that will amplify and reinforce *B* (Hao 2021; Zuboff 2019). A lack of a shared hermeneutical framework could lead to misunderstanding, political instability, and increased polarization that could potentially increase the level of hermeneutical injustice[37] (Fricker 2007) incurred to marginalized groups.

[36] Fricker (2007: 1) defines *testimonial justice* as a situation "when prejudice causes a hearer to give a deflated level of credibility to a speaker's word."

[37] Fricker (2007: 1) defines *hermeneutical injustice* as a situation "when a gap in collective interpretive resources puts someone at an unfair disadvantage when it comes to making sense of their social experiences."

Liars's dividend

The liar's dividend is a phenomenon that occurs in a society moving towards a reality apathy state of affairs, and is first described by Chesney and Citron (2019: 1785). They define it as the situation where liars can deny the truth by pointing out that a video is a deepfake. The 'dividend' is that it pays off for liars to avoid accountability by pointing out that a video is no longer perceptual evidence but has been fabricated. A good example of the liar's dividend has been pointed out by human rights activist Sam Gregory. Gregory (2021) refers to a video where Georgia's former president Mikheil Saakashvili is in the coastal city of Batumi (in Georgia). The ruling party in Georgia invokes the liar's dividend by dismissing the video as a deepfake, but after close inspection the video appears to be real.[38]

3.5 Deepfakes, conspiracy theories and extreme beliefs

In this section the relationship between deepfakes, conspiracy theories and extreme beliefs will be explored. After having defined both terms, I will discuss in the first part how deepfakes can be weaponized in this context and how deepfakes can cause an amplification of harm (Diakopoulos & Johnson 2020; Giansiracusa 2021). In the last part I will analyze this relationship using Baudrillard's simulacra-model.

Definitions

It is hard to come up with a single unified definition for the term 'extreme beliefs' because this is a diverse, ambiguous, value-laden term that, to a great extent, has been historically and culturally defined, and can mean different things in different contexts. A first presupposition that I make, is that extreme beliefs are the beliefs underpinning extremism. A first definition for extremism is the official Dutch government definition: "Extremism – Phenomenon in which ideologically motivated individuals or groups are willing to seriously break the law or engage in activities that undermine the democratic legal order" (NCTV 2021).[39] This definition suggests extreme practices are motivated by ideology, which is "an interrelated set of beliefs that provide a way for people to understand the world" (Cassam 2021: 13).[40] What makes a belief or practice extreme is context dependent and extremism examples in the real world are very diverse (Berger 2018: 24), however, it is possible to identify a set of common features that define extremism. These features should be understood, like fundamentalism (Peels

[38] This event must be put against the backdrop of political tensions in Georgia. Saakashvili was arrested after this video was posted (Lomsadze 2021).

[39] Official definition as found on the website of the National Coordinator for Security and Counterterrorism of the Netherlands. Definition is in Dutch originally and has been translated by the author.

[40] Cassam quotes the authors Uscinski and Parent from their 2014 book *American Conspiracy Theories*.

2021: 224), in terms of family resemblance.[41] Nozick (1997) has identified eight features of extremism that are related to extreme beliefs, extreme actions and extreme groups.[42] Cassam (2021), who builds on the work of Nozick, claims that extremism is not only driven by *what* one believes (ideology), but also *how* one believes and he argues for a psychological *mindset* approach of extremism which not only accounts for beliefs, but also for other psychological characteristics like preoccupations (e.g., purity, victimhood, and humiliation), attitudes (e.g., pro-violence, uncompromising, and intolerance), emotions (e.g., anger, resentment, humiliation and self-pity), and extremist ways of thinking (e.g., apocalyptic or conspiracy). Berger's (2018: 44) extremism definition is about an in-group who can only survive by deploying hostile action against an out-group where the in-group shares the same ideology.[43] I think Berger's and Cassam's accounts provide a good foundation for my working definition for extreme beliefs. Extreme beliefs belong to a group (of people) whose beliefs significantly deviate from a society's social contract to such a degree, that these beliefs are considered to be potentially harmful for society and these people are unwilling to compromise or revise their beliefs.[44] The second presupposition I make for this thesis is that this working definition can be applied to all kinds of beliefs that are considered extreme, like fundamentalist, terrorist and conspiracist beliefs. On this definition it follows those persons who belong to an *echo chamber* may hold extreme views. An echo chamber is a "social epistemic structure from which other relevant voices have been actively excluded and discredited" (Nguyen 2020: 141).[45]

[41] A good definition for fundamentalism understood in terms of family resemblance, can be found in Peels (2020): "A movement is *fundamentalist* if and only if (i) it is reactionary towards modern developments, (ii) it is itself modern, and (iii) it is based on a grand historical narrative. More specifically, a movement is fundamentalist if it exemplifies a large number of the following properties: (i) it is reactionary in its rejection of liberal ethics, science, or technological exploitation, (ii) it is modern in seeking certainty and control, embracing literalism and infallibility about particular scriptures, actively using media and technology, or making universal claims, and (iii) it presents a grand historical narrative in terms of paradise, fall, and redemption, or cosmic dualism."

[42] Nozick's (1997) eight features of extremism, in short: (1) goals and objectives that are on one end of some (political) spectrum, (2) all opponents are viewed as evil, (3) unwillingness to compromise, (4) willingness to use extreme methods, (5) goals must be achieved immediately, (6) organized in groups (no loners), (7) position themselves deliberately as extreme and (8) extremists have a "determinate extremist personality."

[43] Berger (2018: 24) defines an in-group as "a group of people who share an identity, such as religious, racial, or national" where identity is "set of qualities that are understood to make a person or group distinct from other persons or groups."

[44] On this definition it follows that people who hold extreme beliefs would never engage in a reflective equilibrium process, since the pre-requisite for this are the epistemic virtues of open mindedness and willingness to revise one's beliefs and principles.

[45] Nguyen (2020: 141) distinguishes between an *echo chamber* and an *epistemic bubble* which is defined as "An *epistemic bubble* is a social epistemic structure in which other relevant voices have been left out, perhaps accidentally."

A specific subcategory of extreme beliefs is the belief in conspiracy theories. Since conspiracy theory is an umbrella term covering a deep and wide variety of different conspiracies and is subject to interdisciplinary research, it is hard to provide a single, unified definition for this. It is also a value-laden term that has negative connotations associated with it, and it is not clear what the boundaries are if something is a conspiracy or not (Napolitano 2021: 84). Dentith (2014: 3) has tried to provide a neutral, broad, and non-pejorative definition where conspiracy theories are "any explanation of an event that cites a conspiracy as a salient cause." Dentith's account would make everyone a conspiracy theorist and is hence not useful for this ethical thesis. Many academics, coming from different disciplines and backgrounds, have come up with their own definitions each expressing various aspects of conspiracy theories (for examples see e.g., Baurmann and Cohnitz 2021; Cassam 2019; Napolitano 2021; Harambam 2020; Aupers & Harambam 2018). For the context of this thesis, I will provide a working definition that should be understood in terms of family resemblance which is based on the ethical, epistemological, and social aspects of conspiracy theories. Conspiracy theories are theories or beliefs that (i) compete with prevalent scientific knowledge and provide an alternative explanation of complex political, social, medical or natural phenomena[46] (Drążkiewicz Grodzicka & Harambam 2021), (ii) are politically motivated and used as propaganda to promote a political agenda[47] (Cassam 2019a), (iii) are based on *self-insulated* beliefs (Napolitano 2021) meaning that conspiracy theorists are immune to any counter-evidence that is provided in nearby possible worlds, and (iv) conspiracy theories provide a deeper, (quasi-)religious (Cassam 2019a: 60) significance that transcends empirical observation (Aupers & Harambam 2018).[48]

Weaponization

Extreme beliefs and conspiracy theories are often based on one of the three categories of false information (Dobber et al. 2021: 71):[49] mis-information, dis-information or mal-information (see figure 7). Wardle and Derakhshan (2017) provide the following definition for these terms:

[46] Examples of these phenomena are climate change and COVID-19 vaccinations..

[47] It is not always obvious what the real intentions of conspiracy theorists are. A good analytical question to decipher the intentions is the question 'who benefits?' and in many cases the people who produce conspiracy theories are also the people who benefit the most (Cassam 2019a: 34).

[48] Many conspiracy theories have similar elements that can be found in religious beliefs. Conspiracy theories often have a grand narrative about good, evil, suffering and redemption that provides an ultimate meaning and interpretation on how the world works (Aupers & Harambam 2018; De Graaf & Van den Bos 2021; Harambam 2020). A great example of a conspiracy theorist that provides an all-encompassing grand narrative is David Icke. Icke is most well-known for his *reptilian thesis* in which "shapeshifting alien races secretly control our world" (Harambam 2020: 26) and combines this with other narratives, like banking scams, multidimensional universes, and institutional mind control, to a grand super conspiracy theory (see Harambam & Aupers (2019) or Harambam (2020) for an in-depth account and analysis).

[49] Commonly referred to as *fake news* in vernacular.

mis-information, when the news spread is false but no harm is meant
dis-information, when news is false and shared to cause harm
mal-information, when genuine information is spread to cause harm.

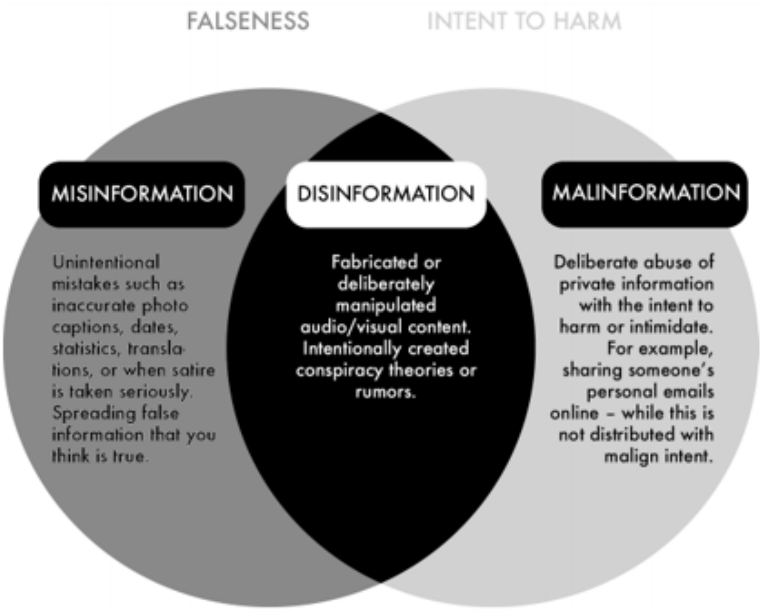

Figure 7 Visual representation of mis-, dis-, and mal-information (source: Van Doorn et al. 2021: 68).

Dutilh Novaes & De Ridder (2021) cogently argue that in many ways there is nothing new in the content and influence of false information, however the Internet and social media have dramatically changed the way how this is being created, disseminated, and consumed. Another study (Vosoughi et al. 2018) found that false information usually travels faster, and to more people, than true information. Most studies are conducted on dissemination of text-based false information; the impact of *visual* false information has been relatively understudied (Dan et al. 2021: 651), but academics expect visual false information like deepfakes, will have a bigger impact than fake news (Rini 2020; Fallis 2020). It is expected that deepfakes will be used to create dis-information (Dobber et al. 2021) and can be weaponized by bad actors (Schick 2020) as a powerful tool to create and distribute content that supports, sustains, and proliferates extreme beliefs and conspiracy theories. However, reality has it that to this date of writing (October 2021), there are not many examples of deepfakes being weaponized in the context of conspiracy theories or extreme beliefs, yet. It is expected this will change in the next five to ten years (e.g., Schick 2020; McGuffie & Newhouse 2020; FBI 2021) because of the increase in speed, scale and ease of use of the technology. In sum, all kinds of technology can be weaponized to create, distribute and consume false information. Deepfakes are not yet weaponized at mass scale, but this is expected to change in the (near) future.

Baudrillard's simulacra-model

A different way to analyze the relationship between extreme beliefs, conspiracy theories and deepfakes is about who gets recognized for holding the epistemic authority for knowledge generation. Because of division of cognitive labor in our society (Baurmann & Cohnitz 2021), we have to trust on experts and institutions for making sound decisions. Traditionally the epistemic power for knowledge production and what is accepted as truth, is held by the government and scientific institutions. Because of various sociological reasons the trust in these institutions has been decreasing over the last few decades (see Harambam 2020 for an in-depth sociological account) and has been contested by conspiracy theorists and people holding extreme beliefs. The central question this poses, is who and what can be trusted as an epistemic authority. As explained above, deepfakes will only contribute and accelerate this process. Baudrillard's simulacra-model (see §3.2) can be a useful model to analyze this. According to Baudrillard (1994), the Western world is moving towards a world that is full of simulations and simulacra where it gets more and more difficult to distinguish what is real and can be trusted. Baudrillard's simulacra-model can applied from a micro-level (for e.g., a deepfake video) to a macro-level (for a society) and explains how our view on reality is mediatized (thus influenced) and distorted and has no reference to empirical truth at all (Harambam 2020: 145). According to Baudrillard (1994: 1) we have arrived at "*the desert of real itself*"[50] which is a situation that is recognizable for conspiracy theorists and people holding extreme beliefs who emphasize "a world where sign and referent, image and reality, truth and fiction are difficult to distinguish" (Harambam 2020: 214). There is one difference between Baudrillard and conspiracy theorists; the first claims there simply is no deeper truth that underlies simulation whereas the latter deploy a hermeneutics of suspicion and want to tear off the mask (Anderson 2019) of the simulacra until they have found the real truth. In short, Baudrillard's simulacra-model helps to analyze how trust and epistemic uncertainty drive a quest for the real truth in a world, to paraphrase Baudrillard, that is moving towards a state of simulation.

3.6 Will empathy save the *Homo Syntheticus*?

After reading this chapter on deepfakes one might get a dystopian feeling that our world is going to collapse under a torrent of deepfakes that are looming. It is good to remember that this chapter predominantly focusses on the potential harm that can be caused by deepfakes, however, to develop a balanced view on the impact of deepfakes it is an epistemic virtue to take a broader perspective and take the positive impact of synthetic media into consideration as well. Based on the current academic literature

[50] Italics in source.

on deepfakes it follows that our society is entering an era in which they have to learn how to deal with deepfakes and the impact this will have on our perception of reality and trust. Van Doorn et al. (2021: 49) have coined this persona the *Homo Syntheticus*, which they define as "a post-reality species that only subjectively perceives reality and where technology alters its relationship with that reality." Their claim is that people have always played with reality throughout history and that technology determines to what extent reality can be manipulated. Their theory, based on theories of play by the historian Johan Huizinga and the sociologist Erving Goffman (Van Doorn et al. 2021: 47),[51] claims all people do is play with reality by projecting an ideal image of ourselves leveraging the technology we have at our disposal. In other words, people are used to subjectively perceive reality and the human species will be able to adapt and evolve to the new synthetic reality that deepfakes will bring about. In the end humans are game-playing storytellers (Van Doorn et al. 2021: 7) using technology to engage with reality.

With the advent of technologies like AI, deepfakes and robots, the anthropological question 'what makes us human' is becoming more and more important. Many social academics like Sherry Turkle (2021), Van Doorn et al. (2021) and Clifford Anderson (2021) believe empathy is what sets humans apart from other artificial entities like robots. Empathy is "allowing us to know what other people are thinking and feeling, to emotionally engage with them, to share their thoughts and feelings, and to care for their well–being" (Stueber 2019). Whether empathy is going to make the *Homo Syntheticus* more human or not, is going to be an important question that will define the synthetic era ahead of us. Whatever is going to happen, it is clear that in the synthetic era humans will be faced with ethical and ontological questions on what it means to be human and what it means to live a good life in times where fake might even seem more real than the truth itself.

[51] Johan Huizinga coined the term *Homo Ludens*, which can be translated as playing man, in his 1938 book *Homo Ludens: A Study of the Play Element of Culture*. The main argument of this book is that play is an essential part of man's culture and man is constantly playing with reality (Van Doorn et al. 2021: 47). Erving Goffman is most famous about his theory where man is an actor on a stage and is constantly playing different roles and thus able to manipulate reality. Goffman wrote this theory in his 1956 book *The Presentation of Self in Everyday Life* and claims that people are more concerned about the impression they leave behind than about finding the truth (Van Doorn et al. 2021: 8).

4. Introduction to Reflective Equilibrium

In this chapter an introduction will be provided to the method of *Reflective Equilibrium*. Reflective equilibrium is one of the most frequently used methods in contemporary moral and political philosophy to systematically analyze and assess moral problems (Arras 2009; Knight 2017; De Maagt 2017). In the first section the details of the reflective equilibrium method and its advantages will be explained and in the subsequent section various criticisms and rebuttals to these will be covered. In the final section the value of this method in the context of this thesis will be discussed.

4.1 Reflective Equilibrium in detail

The term 'reflective equilibrium' has been coined by the American philosopher John Rawls in his book *A Theory of Justice* (1971) in which he used this to develop, advocate and justify his ideas about justice as fairness (Rawls 1971; Cath 2016; Daniels 2016). Broadly speaking, reflective equilibrium is a method which is used to reflect and think about complex, multi-facetted problems by moving back and forth between initial judgments, sometimes referred to as intuitions, and governing principles or beliefs until some form of equilibrium has been reached. The outcome of this process of constantly refining and revising beliefs at all reflective levels is to "seek coherence among the widest possible set of beliefs that are arguably relevant" (Van der Burg et al. 1998: 1) and this can help provide an account of justification for the answer to the question what one ought to do. Reflective equilibrium has a wide variety of applications and is used for finding accounts of justification in both moral and non-moral cases (e.g. Cath 2016; Daniels 1996, 2016; Sheridan 2007; Van der Burg et al. 1998), however in the context of this master thesis the scope is limited to ethical cases only. Two of the major reasons why reflective equilibrium is commonly used in fields like applied ethics and philosophy is its (i) close resemblance to human intuition in navigating moral problems and (ii) its impartiality for justification towards ethical theories or foundational systems of belief like e.g. religion (Arras 2009, Daniels 2016) or as John Rawls put it: "The independence of moral theory from epistemology arises from the fact that the procedure of reflective equilibrium does not assume that there is one correct moral conception" (Rawls 1974: 9). Both Arras (2009) and Daniels (2016) indicate that reflective equilibrium is closely related to the *inductive* scientific method where researchers try to find a law or principle based on a set of empirical observations or data points, they have collected to explain or justify a phenomenon they are researching.

The main goal of reflective equilibrium is *not* to find the truth but to find justification for a position, i.e., the beliefs and judgements, one holds in a moral case. The key idea behind the process of reflective equilibrium is for a thinker to test (Daniels 2016) the beliefs and judgments she holds against other different, or conflicting, beliefs and judgments and revise and refine her beliefs at all levels after appropriate reflection. After the reflective process the thinker's beliefs are justified and she will be in a *state of equilibrium* regarding this moral matter. It should be noted that many proponents of reflective equilibrium indicate that this state of equilibrium is an ideal which most likely will never be accomplished but nonetheless should be strived for so one can come closer to the ideal (Cath 2016). From this it follows that reflective equilibrium is a dynamic process and the state of equilibrium is not fixed, and any new input can cause the thinker to go through the reflective equilibrium process again and revise and refine her beliefs. In the original conception of reflective equilibrium John Rawls thought the reflective equilibrium process to be conducted by one single person only, who I called the thinker (Rawls 1971: 50; Van Thiel et al. 2010), however in applied ethics there are many use cases where reflective equilibrium is used as a method involving many stakeholders (e.g., Arras 2009; Doorn 2010, 2012; Doorn et al. 2018; Griffin 1993; Knight 2017; Van den Hoven 1997). In this thesis the reflective equilibrium process will be conducted from a single person perspective in which I am taking on the role of the thinker.

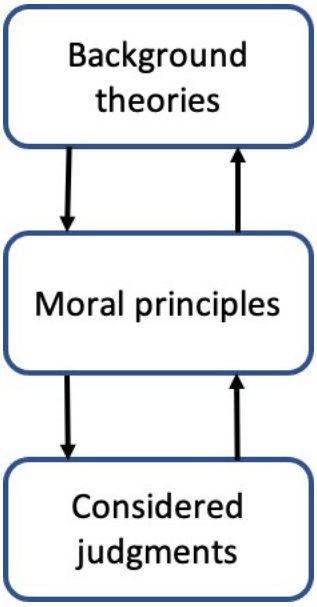

Figure 8. Reflective levels in reflective equilibrium

The reflective equilibrium process takes place by moving back and forth between the following three reflective levels: considered judgements, moral principles, and background theories, as is schematized in figure 2. Knight (2017) points out that the starting point for the reflective equilibrium process is the judgments on a moral case.

According to John Rawls not just any judgment or intuition can be used, but only *considered* judgments, which he defines as "those judgments in which our moral capacities are most likely to be displayed without distortion" (Rawls 1971: 47). Most authors agree with Rawls that the judgments should be held with confidence, in other words, to put a *confidence* constraint on the judgments (Van Doorn et al. 2018; Knight 2017). Another proposed idea would be to put an *epistemic* constraint on the judgments which allow "only justified or warranted judgments, or (more minimally) only those that lack errors" (Knight 2017: 3)[52] to be used in reflective equilibrium. Doorn et al. (2018) argue that putting too much emphasis on what judgments should be considered credible may render the method too exclusive and some judgments are not taken into consideration. They point out that most philosophers are in favour of a credibility threshold whereas most social scholars tend to focus on the inclusiveness of the method.[53] Knight (2017) forcefully argues that both the epistemic and the confidence constraint should be discarded, and to him considered judgments are defined under "conditions favorable for deliberation and judgment in general" (Knight 2017: 3). I agree with Knight that no *a priori* constraints should be put on considered judgments and that the focus should be on inclusivity, in other words, that all voices should be heard, and all judgments are taken into consideration when the reflective equilibrium process is conducted. This does not mean that 'anything goes' and there should be an entry level for judgments, beliefs, principles and participants who participate in reflective equilibrium should that are based on *epistemic virtues*. In his paragraph *How to use the method of reflective equilibrium* Knight (2017: §5) brings the following requirements to bear for the thinker when bringing considered judgments into the reflective equilibrium process:

- No upset, fright, tiredness, or intoxication.
- No conflicts of interest.
- Minimal epistemic competency about the moral topic.[54]
- Open-mindedness and willingness to alter one's views.

The next reflective level in the reflective equilibrium process are *moral principles* that have an impact on the moral case at hand. Moral principles are a fundamental moral building block that govern moral action or as Beauchamp (2016: 81) puts it "A principle is an essential norm in a system of thought or belief, forming a basis of moral

[52] The page numbers for this article deviate from the page numbers of the journal it appeared in. I have used the version that is available on https://doi.org/10.1017/9781316162576.005.

[53] The problem of 'initial credibility' also applies to the other reflective levels (moral principles and background theories) as Doorn et al. (2018: 510) argue in footnote 11 of their article.

[54] Knight (2017: 10) argues that to be willing to reach the correct decision a minimal amount of competency is needed about the moral topic handled in the reflective equilibrium process. Theoretically one should be able to reach a reflective equilibrium without any knowledge but according to Knight (2017: 10) this is "unlikely to have much epistemic value."

reasoning in that system." Verweij (1998: 39) claims that principles have "strong normative authority for the people who endorse them" in other words, principles justify moral judgments and drive moral behavior. An ongoing meta-ethical debate is taking place as to whether principles are fixed and foundational and should leave no room for exceptions or that principles are more like guidelines and can be revised, recontextualized or derived from various analog cases (Beauchamp 2016; Fieser 2021). Principles are often defined at a high level to be as general and substantive as possible whilst they still remain normative. After reading the reflective equilibrium literature I was surprised this mainly focuses on principles and barely mentions (moral) values. If one follows Craig & Moreland's definition of values being "the adherence to some moral proposition that *prescribes* what morally *ought* to be" (2017: 425) then it follows that values overlap the definition of principles for a great deal. In the ethical literature, however, values can be experienced as subjective, being a personal stance or not being a fact, or objective where the values can be derived from an independent source (e.g., religion) outside a person (Craig & Moreland 2017; Blackburn 2008). Van der Burg (1998: 94) considers values to be *ideals* that are not fully realized yet and that "partly transcend contingent, historical formulations and implementations in terms of rules and principles." In short, for this thesis principles govern moral decision making in a current state of affairs whereas values are the ideals to strive for in a future state of affairs. I would argue that values are embedded in the reflective levels of considered judgments and background theories.[55]

In the original proposition of reflective equilibrium, as proposed by John Rawls in *A Theory of Justice*, coherence is sought between considered judgements and moral principles to achieve the status of equilibrium. This is commonly referred to as *narrow* reflective equilibrium (e.g., Cath 2016; Daniels 2016; Doorn 2010) however, most philosophers and ethicists prefer to include the reflective level of *background theories* in the process, which is called *wide* reflective equilibrium (Cath 2016).[56] Background theories are beliefs or theories that influence and/ or complement one's considered judgments and the moral principles that govern these judgements and can be e.g., alternative moral theories, beliefs about psychology, metaphysics (Cath 2016) or social theory (Knight 2017). Proponents of wide reflective equilibrium like Daniels (1996: 22-23) claim that adding background theories to the reflective equilibrium process will increase the epistemic status of the justification of the beliefs held (Van den Beld 1998: 80). Adding background theories is not unproblematic and this view has received its share of criticism like e.g. the arbitrariness of what background theories should or

[55] I disagree with Van der Burg (1998: 94) who argues that values (or ideals as he calls them) should be added as a fourth reflective level or category to the reflective equilibrium method.

[56] In this thesis whenever I refer to reflective equilibrium, this is wide reflective equilibrium and not narrow reflective equilibrium, unless specified otherwise.

should not be added (Kelly & McGrath 2010). However, in applied ethics the vast majority of reflective equilibrium practitioners prefer to use wide reflective equilibrium and therefore this method will also be used in this thesis.

After collecting all the data, the process of reflection starts where the thinker tries to achieve the state of equilibrium. During this process the thinker reflects on what the impact of the principles and background theories is on the considered judgments and whether these need to be refined or revised when there are e.g. principles that conflict with her judgments.

Out of the many authors that have written about 'reflective equilibrium' that are listed in the bibliography, Knight (2017: §5) has provided the most practical recipe how to conduct a reflective equilibrium process. He describes a seven-step process which is listed in table 1 below.

Steps in the reflective equilibrium process	**Description**
1. Making considered judgments.	Being sure that the judgments meet "Rawlsian 'conditions favorable for deliberation and judgment in general'" (Knight 2017: 11).
2. Make a list of contending moral principles.	Draw up a list of all the potential contending principles that potentially play a role in the case.
3. Testing the judgments and principles against each other.	For each principle the thinker reflects on the impact it has on the considered judgments. The outcome is either to accept the principle, to reject it or to revise your judgment.
4. Introduction of background theories.	Introducing relevant background theories for reflection.
5. Review the process.	Careful reflection using all the components gathered in the process so far.
6. Establishing priority rules.	In case of conflicting principles, the thinker needs to consider which principles are prioritized.

7. Conclusion of the reflective equilibrium process.	Conclude whether a reflective equilibrium is found between the judgments and principles or not.

Table 1 The seven steps in the reflective equilibrium method (Knight 2017: §5).

4.2 Criticisms

After researching and working with the reflective equilibrium method as an ethical tool, I have experienced both the advantages and the drawbacks of it. Reflective equilibrium, being used extensively as a method in moral philosophy, has also received its share of criticism in the academic literature (Knight 2017). In this section I will share four major criticisms (and rebuttals to these criticisms) that I have experienced myself and which are also described in the academic literature. These criticisms are based on practicality, arbitrariness, unreasonable beliefs, and conservatism.

Practicality

I would have expected with reflective equilibrium being the most widely used method in applied ethics, that I would be able to find a lot of case studies in the academic literature that describe how this method is being used on concrete cases in practice. The vast majority of scholarly articles do a great job advocating why the method works better than other ethical methods (e.g., Rawls 1974; Griffin 1993; Daniels 1996, 2016; Van den Hoven 1997), describing the drawbacks of it (e.g. Kelly & McGrath 2010) or discuss to what extent reflective equilibrium can provide normativity (e.g. De Maagt 2016) but they lack good descriptions of case studies. In some articles that claim to describe a specific ethical case (e.g., Sheridan 2007) using reflective equilibrium the largest part of the text discusses *what* the method entails rather than explaining *how* it can be applied. Doorn (2010, 2012), Doorn & Taebi (2018) and Schroten (1998) e.g., do provide practical case studies that describe how reflective equilibrium is used however, these case studies are applied to a multi-stakeholder cases instead of taking a single-thinker perspective as I am doing for this thesis. In short, there is an extensive body of literature written about reflective equilibrium but this has not been very helpful in my research from a practicality perspective.

Arbitrariness

The positionality of the participants who are involved in a reflective equilibrium process determines to a great extent the outcome of the process. For my thesis I am aware of my own positionality and the accountability I need to provide for this. This criticism that reflective equilibrium can lead to moral arbitrary outcomes can also be found in the literature as Cath (2016: 221) writes “different people may have very

different initial beliefs and, hence, might reach different equilibria when they apply this method." Authors have rebutted this criticism by claiming that reflective equilibrium does not deny the existence of moral pluralism and is also not a 'magic bullet' that can lead to a unified moral judgment. Instead, the method is specifically designed to deal with multiple, and often conflicting perspectives and treats every moral input with equal epistemic status. In the interpretation of Rawls (1974: 9) it follows that if people don't converge after having applied the method of reflective equilibrium than there are no objective moral truths. Ultimately reflective equilibrium is not designed as method to find the truth but as a method to provide justification for one's beliefs. I agree with these rebutting authors that pragmatism and inclusivity outweigh the potential harm of arbitrariness. The thinker (who conducts or guides the reflective equilibrium process) needs to be vigilant for her own positionality and needs to be accountable for that.

Unreasonable beliefs

Another widely documented criticism is that reflective equilibrium is too dependent on the quality of the considered judgments (e.g., Knight 2017; Cath 2016; Daniels 2016) but has been mostly advocated by Kelly and McGrath (2010). They claim that if you start reflective equilibrium with unreasonable, implausible, or repugnant considered judgments than "it turns out that impeccably following that method could lead one to views that are *unreasonable*" (Kelly & McGrath 2010: 346). Knight (2017: 6-8) disagrees with this objection as he points out that Kelly and McGrath implicitly assume that from flawed moral principles automatically flawed considered judgments follow. According to him this may be true for narrow reflective equilibrium, but it does not apply to wide reflective equilibrium because many conflicting background theories and principles are taken into the reflective equilibrium process which has a self-healing effect to expunge judgments that are unreasonable. In addition, Knight claims that neither reflective equilibrium nor the scientific method are "*guaranteed* to rid people of unreasonable beliefs. But that doesn't change the fact that both are more likely than alternatives to provide individuals with reasonable beliefs, by exposing them to the most compelling evidence that is available in their respective fields" (Knight 2017: 8). I agree with Knight this objection doesn't hold but it is important that reflective equilibrium practitioners are aware of this potential pitfall. As a researcher and thinker I shouldn't be too rigorous in applying the method but I need to always assess the outcomes in the appropriate context.

Conservatism

Philosophers like Singer, Brandt and Hare have pointed out that reflective equilibrium is being too conservative by putting too much weight on conforming the moral principles to moral judgments (Knight 2017; Cath 2016). This is supported by the

claim that moral judgments are derived from untrustworthy sources like "discarded religious systems" (Cath 2016: 221), have been propagated through our genes and elicit an emotional rather than a rational response that is informed by reason (Knight 2017: 8). The rebuttal from reflective equilibrium proponents is to take in these views as background theory and in case of a conflict between beliefs and judgments that an agent will be able to revise and refine her judgments and her initial beliefs. In addition, Knight (2017: 10) argues that because of bringing in these background theories in the reflective equilibrium process, this might shift the weight of personally related judgments to more universal judgments. Refining judgments is a proper part of what the ethicist David Brink calls a "dynamic dialectical process" (Knight 2017: 10). I think this objection underscores why it is important for the thinker and participants in the reflective equilibrium process to have the epistemic virtue of open-mindedness when participating.

4.3 Reflective Equilibrium in the technological context of this thesis

Reflective equilibrium is a method that is commonly used in ethical cases that are related to technology (e.g., Doorn 2010, 2012; Doorn et al. 2018; Schroten 1998; Sheridan 2007; Van den Hoven 1997). Van den Hoven (1997) argues that in a time where technological concepts like deepfakes are rapidly changing, it is essential to revisit and reformulate moral principles and judgments that are related to this. In his article he argues that wide reflective equilibrium offers the best model for this and "incorporates the best of the universalist and particularist worlds" (Van den Hoven 1997: 242). The first position, universalism, is a top-down approach where moral principles govern moral technological cases. This position, what Van den Hoven (2017: 236) dubs as the "engineering model" for ethics is a common ethical approach in the world of technology. The major objection to this is that the models do not work well with edge cases and exceptions and is too rigid to use in a fast-changing environment. The other position described by Van den Hoven (1997: 240) is particularism, where moral judgments are based on what Aristotle calls *phronesis*, practical wisdom based on discussing individual cases. In bioethics, which has a lot of similarities with technology ethics, this approach is called casuistry (Arras 2009; Beauchamp 1996). Van den Hoven's (1997: 241) major objection against this method is that it "black-boxes moral justification" when similar cases are judged there is no reference to a moral principle. Wide reflective equilibrium is a good ethical model to use in technology related ethical cases that are subject to change, like the topic discussed in this thesis. Van den Hoven also emphasizes that using the wide reflective equilibrium method is perhaps the best method to use from an epistemic virtue point of view: it uses no epistemically privileged propositions and incorporates a "doctrine of intellectual

responsibility" (Van den Hoven 2017: 243) for agents involved in reflective equilibrium to strive for open mindedness and reduction of failures.

4.4 Conclusion

The philosopher T.M. Scanlon claims that reflective equilibrium, if properly applied, probably is "the best way of making up one's mind about moral matters" (quoted in Cath 2016: 216) and I agree with him, based on the many successful accounts in which reflective equilibrium has been applied in moral matters. Reflective equilibrium is a proven and robust method and in addition to that, it is a method that is closely related to hermeneutics, as it forces the researcher to expose her own initial beliefs on a certain matter. This makes the researcher aware of her own position and embeddedness in the moral matter. In the next steps the researcher comes up with theories that account for her initial beliefs and engages in a reflective process if there are conflicts between the beliefs and the theories, just if she has reached a "state of reflective equilibrium" (Cath 2016: 214). Through a robust method of reflection, the researcher is challenged to take other perspectives into consideration and account for this. Reflective equilibrium is not a 'one size fits all' method and can be applied in many ways and many alternative reflective equilibrium methods have been developed (e.g. Van Thiel & Van Delden 2010; De Maagt 2017; Van der Burg 1998) which makes it necessary for the researchers to assess which what steps are needed to conduct the reflective equilibrium process. For this thesis I am using a simplified version of the process described in Knight (2017: §5) in which the process is being conducted from a first-person perspective in which I take on the role of the thinker. The process will be briefly explained in chapter 5 *Cases*. In short, reflective equilibrium is a suitable method for this thesis because it helps to come as close as possible to a coherent moral framework for the use of deepfakes in context.

5. Case studies

I will apply the reflective equilibrium method as described above to three selected cases which will be outlined in this chapter. This chapter starts off by providing a justification for the selected cases, followed by a brief description of each case. The application of reflective equilibrium to these cases will be covered in the subsequent chapters.

5.1 Justification for case study selection

For this thesis I have selected three case studies. These case studies do not necessarily represent a larger population of other case studies, but the lessons learned from these cases can be applied to ethical cases in different contexts. Each case reflects different moral and technological aspects on how deepfake technology impacts what one believes and holds to be true from an epistemological perspective. Two case studies are real life cases that have taken place in the last two years. The last case study is a thought experiment about what might occur ten years from now when deepfake technology has matured. The cases have an increasing level of technological sophistication to focus on the impact technology has on the moral properties of the case. A narrow timeframe has been selected because the focus of this thesis is on deepfake technology which did not exist as such ten years ago and the development of this technology is taking place at an exponential pace (Chesney & Citron 2019: 1753; Dorubantu 2020: 187).

The selected cases are based on the following criteria: (i) deepfake or alleged deepfake technology[57] is used, (ii) the case is related to conspiracy theories or to a minority view that conflicts with the mainstream view, (iii) the case has taken place in a society that holds a western worldview, preferably in Europe and (iv) the case deals with stakeholders who hold different conflicting moral perspectives on the case. The common thread across the various cases is the increasing level of technological sophistication to influence reality and a decreasing level of trust in the epistemic truth value of media. Each selected case represents a different level in Baudrillard's (1994: 6) simulacra model.[58] In short, Baudrillard's model discusses a world where meaning depicted in images increasingly gets separated from reality (Morris 2021: 322) until the situation has become a simulacrum in which there is no relation to reality at all. The first case refers to the second phase, perversion of reality, in which reality is

[57] It must be noted that for the first case, deepfakes were initially assumed to have been used but it turned out later to be a cheapfake.
[58] This model is being discussed in more detail in chapter 3 *Deepfakes defined*.

obscured and denatured by an alleged deepfake that turned out to be a cheapfake; the second case is a case study in which a deepfake movie is used that masks the absence of reality (third phase in Baudrillard's model). The third case study is a fictitious case study based on a thought experiment, that takes place in the future; This case study is a simulacrum, which is the fourth phase in the model.

5.2 Case 1. Alleged Deepfakes

On 21 April 2021, members of the Dutch parliament Foreign Affairs committee[59] were supposed to have a Zoom video call with Leonid Volkov, who is the spokesperson for the incarcerated Russian opposition leader Alexei Navalny. After some time in the meeting, the video call turned into such a weird conversation that the parliamentarians initially thought they had encountered a deepfake version of Volkov (Verhagen 2021). Members of parliament from other countries like Estonia, Latvia, Lithuania and the UK have had similar experiences with the fake Volkov (Roth 2021) and the joint Baltic Foreign Affairs Committee chairpersons even put out a joint statement in which they warned against the use of deepfakes by Russia in spreading disinformation (Brouwers 2021).[60] However, a few weeks later it turned out Leonid Volkov was *not* a deepfake, but he was a famous Russian comedian who had impersonated Volkov (see figure 9). The Russian comedians released a YouTube-video in which they declared that they had pulled a prank with the Dutch democratic representatives (Van Assen 2021; Vovan222prank 2021).[61] Despite this being a cheap trick, Dutch members of parliament were not amused, and it has impacted the way they perceive reality in the future. According to Kati Piri, representative for the Dutch Labor Party, this has been a wakeup call for herself or to quote her "we have to be much more aware of these situations in the future; in the modern world there's more possible than one can imagine" (Van Assen 2021).[62]

[59] An overview, member list and job description of this committee can be found on the website of Dutch Parliament https://www.tweedekamer.nl/kamerleden_en_commissies/commissies/buza.
[60] The full statement can be read in the tweet sent by the Latvian chairperson RIhards Kols https://twitter.com/RihardsKols/status/1385576305056534533.
[61] *Vovan222prank* is the name of the account that posted the video on YouTube. Since I don't know the real name of the person who uploaded or created the video, I am using the account name as a reference.
[62] Quote appeared originally in Dutch and has been translated to English by the author.

Figure 9 Screenshot of the fake Leonid Volkov (source: Vovan222prank 2021).

The reason for selecting this case is that is a good representation of the public opinion around deepfakes in 2021. This video was initially qualified as a deepfake but after a few weeks turned out to be a cheapfake. This case study exemplifies that in 2021 cheapfake technology can be as impactful as deepfake technology (Harris 2021). The case not only made headlines in the Netherlands but also in other countries (Roth 2021) and put the spotlights on the potential dangers and advantages of deepfakes. The incident in itself was considered innocent, but it exposed how potentially powerful deepfake technology can be if even members of the parliament, in the heart of a democracy, can be fooled. The interesting paradox in this case is the discussion to what extent the freedom of speech should be limited or not. The members of Dutch parliament were talking to a representative of Navalny, the person whose freedom of speech has been severely limited because he holds a different view than the Russian government, whereas the same members of parliament were talking to a Russian comedian who exercised his freedom of speech. The moral properties of this case revolve predominantly around to what extent the freedom of speech can be exercised and in what context a person needs to be accountable for using fake technology, defined in the broadest possible terms, and to what extent minority voices are being suppressed. It can be argued that Navalny is suffering from testimonial injustice (Fricker 2007; Moyaert 2019). Testimonial injustice takes place “when prejudice

causes a hearer to give a deflated level of credibility to a speaker's word" (Fricker (2007: 1) which applies to Navalny in two ways: (1) because he is incarcerated, he is unable to voice his opinions and (2) the Russian authorities are decreasing his credibility as a speaker and a knower. The impersonated spokesperson, Leonid Volkov, is suffering from what Rini and Cohen (2021) call *illocutionary wronging*, which is an undesired speech-act. Volkov has been wronged in his capacity as a speaker because the Russian comedians who impersonated him, were saying things he would never have said himself. Rini and Cohen describe illocutionary wronging in the context of potential harm of deepfakes and it is remarkable that a cheapfake, being an impersonation by the Russian comedians, can have a similar impact.

The Russian comedians clearly think this will we a news parody whereas the Dutch members of parliament view it as manipulation. The central ethical issue underpinning this case study, is to morally assess the impersonation of a political person or his representative against a tense political backdrop. In sum, this case study is about manipulating reality using cheapfake technology against the backdrop of a broader geopolitical and societal context.

5.3 Case 2. Deepfakes today

The second case took place in 2020 in Belgium where the environmental activist group *Extinction Rebellion Belgium* posted a video (see figure 10) on their website in which the Belgian prime minister at that time Sophie Wilmès, delivers a speech in which she links the COVID-19 virus with the climate crisis (Galindo 2020; Langguth 2021). Wilmès addresses the Belgian nation that there is an urgent need to tackle the climate crisis and pandemics are one of the consequences of the deep climate crisis we are in. The video is being accompanied by the hashtag #TellTheTruthBelgium (Langguth 2021; Extinction Rebellion 2020a) and turns out to be a deepfake video of the prime minister based on a previous speech she delivered.

Figure 10 Screenshot of deepfake video of Belgian prime minister Sophie Wilmès (source: Extinction Rebellion 2020a).

It is not hard to detect this video as a deepfake since the authors intentionally show this near the end the video (see figure 11). The deepfake video has been created by an Extinction Rebellion (XR) volunteer (De Standaard 2020) and clearly demonstrates that no marketing agency, sophisticated IT skills or visual special effects are required to make a deepfake.

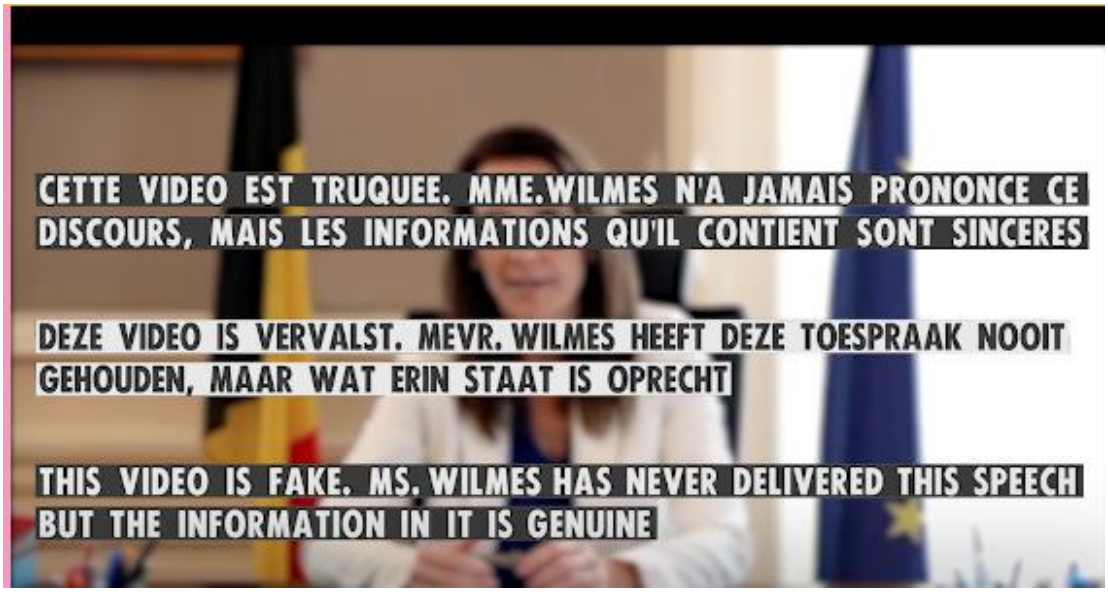

Figure 11 Screenshot of labels indicating the video of Wilmès's speech is fake (source: Extinction Rebellion 2020a).

The deepfake is part of a series of activities XR organized during the COVID-19 lockdown to keep pressure on politicians and leaders to raise awareness for the devastating impact of the climate crisis (Galindo 2020). XR intended to use this video to provoke the Belgian people to a conversation about climate change and according

to a XR Belgium spokesperson this campaign has broken all the records (Holubowicz 2020) and this deepfake generated 100,000+ views (Extinction Rebellion 2020b). XR did not intend to do harm or fool anyone, however, there have been commentators and people who were confused by this video and took the video for real (Holubowicz 2020). XR justified the use of deepfakes by clearly labeling the video as fake and warranted the fabricated content delivered by Wilmès by publishing the speech verbatim on their website with extensive footnotes referring to scientific publications (Extinction Rebellion 2020a).

The interesting perspective this case study offers, is to explore the impact of using reality altering technology like deepfakes by political organizations or pressure groups. There are other examples of political deepfakes, however this video is considered to be the *first* adverse use in which a political figure was targeted, and her identity was misappropriated (Ray 2021: 986). Are deepfakes another tool in the toolbox of marketing and PR professionals, a new way to create satire, or will the impact be much bigger if this technology is being weaponized by minority groups or extremist groups to deliberately change reality. In this case study I'll explore the ethical impact of how deepfakes can be deployed by political actors, pressure groups and other organizations to voice their opinion and to engage with the audience. This case study shows an example how deepfakes can be deployed with good intentions, but its epistemic consequences are similar than when it's intentionally used to create disinformation by e.g., actors holding extreme beliefs and conspiracy theories. In addition, this case study is a good example of the deepfake state of affairs in 2021, in other words, how deepfakes are currently being used, perceived and deployed in our Western society. The central ethical issue underpinning this case study, is what it means from a moral and epistemological perspective, when deepfakes are being weaponized by political actors and what potential harm could this bring about?

Political actors, being defined as all actors that have a political agenda they want to pursue, use deepfakes for propaganda purposes. In this process the various stakeholders hold different conflicting interests and the ethical issue that will be analyzed in the reflective equilibrium process is to morally assess if political actors are justified to use deepfakes, in order to pursue their political goals, and who is responsible for the harm being done.

5.4 Case 3. Deepfakes ten years from now

The third case study is a description of a potential event that could take place ten years in the future, in other words, the content of the case study is being informed by a thought experiment. To make an informed decision about the content of this thought

experiment, I have conducted interviews with several deepfake experts and have asked them how they see deepfake develop and which role this will play in our society ten years from now. In addition, I have consulted several articles and books that make projections about the future of deepfakes (e.g., Schick 2020; Meckel & Steinacker 2021; Rini 2020; Diakopoulos & Johnson 2020; Vaccari & Chadwick 2020; Dan et al. 2021) and I have consulted recent academic articles that describe COVID-19 and anti-vaccination conspiracy theories (e.g., Bortolotti & Ichino 2020; Pierre 2020; Jutzi et al. 2020; Walter et al. 2020).

The scenario of this case study is about the impact of deepfakes in 2031. I presuppose that deepfake technology has progressed to a level that anyone with a smartphone and an internet connection is able to create a high quality, high resolution deepfake video or audio. The generated synthetic content is of such a quality that it will be impossible to make a clear distinction between what is real and what is fake (Schick 2020). This will dramatically impact the way we perceive reality and how our cognitive and epistemic faculties are changing from a *hermeneutics of faith* to a *hermeneutics of suspicion* (Anderson 2019). The fictitious case that I constructed is about a group of conspiracy theorists who are living in a world that is dealing with the aftermath of the COVID-19 pandemic. Since the start of the pandemic the amount of people who believe in conspiracy theories for a causal explanation of this pandemic has increased (Bartolotti & Ichino 2020; Jutzi et al. 2020). There was already a growing group of conspiracy theorists (also known as *anti-vaxxers*) who believe that governments and major pharmaceutical companies (Big Pharma) are promoting vaccinations while these may cause considerable harm like e.g., autism (Pierre 2020). The COVID-19 pandemic and its vaccination strategy worked as a catalyst for these anti-vaxxers (Ashton 2021). There is prolific vaccine mis- and disinformation to be found on the internet, which is created and distributed by anti-vax groups, but also Internet bots (Pierre 2020), Russian internet trolls (Walter et al. 2020) and the algorithms of social media platforms play an important role in this.

The fictitious case study takes place in the Netherlands in 2031. The Dutch population is still dealing with the aftermath of the COVID-19 pandemic. In order to live in this society, the European Union (EU) has mandated that everyone should have a digital COVID certificate so they can travel, work or visit places like bars and restaurants. In order to maintain this certificate, one needs to get an annual repeat vaccination. There is still a lot of resistance against this government policy and the amount of mis- and disinformation and subscription to conspiracy theories is still an important phenomenon that informs the public discourse around the need for repeat vaccinations. One of the leading influencers who are against annual COVID-19 repeat vaccinations is a group called *No Vaccine Repeat* (NVR). NVR considers repeat

vaccinations to be a plot by the EU and major pharmaceutical companies (Big Pharma) in order to control and manipulate the European population. The leader of NVR, called Bill d'Angelo[63], is a charismatic person who is a prolific author and creator of content in which he advocates against the use of repeat vaccines. He claims that anyone who disagrees with him are *sheeple*, a portmanteau of sheep and people. NVR has a large base of followers and people who sympathize with their ideas. NVR is inundating the internet with all kinds of media (think e.g., movies, podcasts, blogs, comments to news articles etc.) that supports their statements. The Dutch government considers NVR to be a group holding extreme views and like many of these groups, NVR has embraced the use of media and technology (Peels 2020; De Graaf 2021: ch.4) to promote their message. They have embraced synthetic media technology and have a dedicated group of volunteers who churn out a flurry of deepfakes on an ongoing basis. The majority of these deepfakes are detected and labelled as deepfake by the mandatory deepfake detection algorithms that the internet providers and social media platforms use. For this case study I presuppose that NVR has created a deepfake that is a *false positive*[64] and is classified as true, verified information by the detection algorithms. This deepfake video is a video in which prime minister Mark Rutte[65] is having a private conversation with Maurits Majoor[66] who is the CEO of a large pharmaceutical company in the Netherlands that creates COVID-19 repeat vaccines. In this video Majoor suggests to Rutte that NVR should be declared an illegal criminal organization because it is holding extreme views that undermine the trust in democracy and science. People affiliated with NVR are incited to take actions that go way beyond non-violent civil disobedience and Majoor even has clues that NVR is planning to commit arson in one of their warehouses where vaccines are stored. He argues that Bill d'Angelo should be arrested and incarcerated and that anyone who is affiliated with NVR should be targeted as potential terrorists who should be closely monitored. This video goes viral, and NVR is using its momentum to distribute their political message that the EU and Big Pharma use the repeat vaccines to control the Dutch population. This Dutch population has been exposed to so many deepfakes over the last few years that they have learned and been trained and coached, to trust the deepfake detection algorithms over their own cognitive faculties. Rutte and Majoor are claiming this video to be a clear deepfake (liar's dividend), however, the deepfake video causes a lot of confusion and research shows that trust in the government and repeat vaccines has diminished.

[63] This is a fictitious name and is only used to make this case study more tangible.
[64] A *true positive* in this scenario is a deepfake that has been detected and classified as a deepfake by the deepfake detection algorithms.
[65] I presuppose that Mark Rutte is still prime minister of the Dutch government in 2031 and this is his nineth term.
[66] This is a fictitious name and is only used to make this case study more tangible.

This case study takes place in a society that is suffering from an infocalypse (Schick 2020) in which one of the survival mechanisms is an increased level of trust in detection algorithms. If these algorithms classify the video as true than it can be safely assumed this video is true. This mechanism has led to a society that is still able to operate on a hermeneutics of faith but also to a continuous cat and mouse game between the creators of deepfakes and the creators of detection algorithms (e.g., Giansiracusa 2021; Kerner & Risse 2021: 86; Yadlin-Segal & Oppenheim 2021).[67] The ethical issue underpinning this case study is to assess the moral implications of a deepfake that has been created and distributed by an organization holding extreme beliefs and conspiracy theories in a world where the population is mainly trusting in technology to assess the truth.

5.5 Reflective Equilibrium method applied

In the next chapters the reflective equilibrium method will be applied to the cases described above. The method used in this thesis is a simplified version of the seven-step reflective equilibrium method that is described in §4.1, table 1. It starts off with an overview of the ethical issue, the stakeholders and their interests. Subsequently followed by an account of the reflective equilibrium process which consists of three components: (i) an overview of the relevant considered judgments, (ii) principles and relevant background theories that are at play and (iii) the process of reflection. For each component I shall briefly outline below how I collected and processed the data and did the reflections. The chapter will be closed with a summary description of the outcome of the ethical considerations and reflections. As a reminder, the reflective equilibrium process will be done from a first-person perspective in which I am taking on the role of the thinker.

Considered judgments

For each case study I have collected the considered judgments based on the provided case description above and the (public) sources that I used for each case. If judgments are not clearly described in the source, or are implicitly present than I have complemented, augmented, and extrapolated these judgments based on the best of my knowledge. Being the thinker. this is justified because this helps to provide a better overview of the moral landscape, the overview of all relevant moral properties and

[67] For this thesis I presuppose that deepfake detection algorithms can keep up with the expected increase in volume of deepfakes and the algorithm's accuracy provides sufficient evidence to believe their classification. This is contrary to what some authors think will happen, as they expect it will become impossible to train deepfake detection algorithms because of the sheer volume of deepfakes (see e.g., Schick 2020: 134-145 for a discussion).

judgments, of each case. If a judgment does not have a reference attached, it can be assumed that I have augmented the judgment.

Principles and background theories

The next step in the reflective equilibrium process is to generalize for each considered judgment what moral principles and background theories might explain the judgment. In moral cases almost always conflicting principles and background theories arise. I will infer the principles and background theories based on the case description and available sources. It can be assumed that all principles and background theories are indirectly derived from the sources unless this has been specifically referenced. If the latter is the case than a reference to the source will be added.

Reflection

After collecting and preparing all the moral data needed, the process of reflecting and considering can start. The thinker will review each conflicting judgment and principle/ background theory and revise the judgment or principle. After completing this process, the thinker has reached a state of equilibrium for this case, in other words, the thinker has found a coherence between the thinker's position and the moral principles at stake in the case studfy. It should be noted that the state of equilibrium applies for the thinker and may and cannot be generalized to a normative judgment or statement, which is also *not* the goal of this thesis. The reader will get an overview of the considered judgments and principles that are at stake for each case study and can use the thinker's considerations and reflections as input for further applications like e.g., academic research or applied ethics.

6. Case study 1. Fooled by Fakes

6.1 Ethical issue

In short, case study 1 is about the Dutch Parliament Foreign Affairs committee who think they talked to Leonid Volkov, the spokesperson of Navalny, and thought they were fooled by a deepfake, but instead they were pranked by a cheapfake being Russian comedians who impersonated Volkov. The central ethical issue underpinning this case study is to morally assess the impersonation of a political person or his representative against a tense geo-political backdrop.

6.2 Stakeholders and interests

Table 2 below contains a schematic overview of the relevant stakeholders, their role and their interests that are part of case study 1.

Stakeholder	**Role and Interests**
Leonid Volkov	Chief of staff of Alexei Navalny and represents Navalny and his interests to the outside world. He is being impersonated by the Russian comedians.
Alexei Navalny	Russian opposition leader and anti-corruption activist who is currently being detained in Russia. He is serving time for parole violations and has been charged with "creating an organisation that 'infringes on the personality and rights of citizens'" (Balmforth & Zverev 2021). His interests are fighting corruption and a better, more transparent democracy in Russia. He is being supported by many EU governments. Nalvany is not an active stakeholder in this case, but his interests form the origin and geo-political backdrop for this case.
Dutch Parliament Foreign Affairs committee	Fixed committee of members of Dutch parliament who discuss foreign affairs matters. Their interest was to hear from Volkov how the Dutch parliament could help Navalny progress his interests. They work from a default epistemic state of trust that the people they are talking to are real and not fabricated or impersonated by someone else.
Russian comedians	Their role is to provide a satirical view of the world. Their interest is to exercise their freedom of speech and their right to entertain their audience.

Russian government	Is indirectly involved in this case as they have incarcerated Navalny. This makes them responsible for Navalny's testimonial inability to communicate with the outside world.
Political influencers	A good example is the former member of the EU parliament and Stanford University research fellow Marietje Schaake. She thinks deepfake technology can disrupt democracy and should not be allowed to be used in political campaigns (Van Assen 2021; Schaake 2021).

Table 2 Stakeholders and interests of case study 1.

6.3 Considered judgments

All relevant considered judgments can be found below, grouped by stakeholder.

Leonid Volkov

The real Leonid Volkov stated in an interview that he most likely has been impersonated by the Russian pranksters Vokan and Lexus which has happened a couple of times before (Moscow Times 2021).[68] He continues in the interview "We call them pranksters, but in reality, they are well paid employees of the Russian government, it's more serious than it seems" (NOS Nieuws 2021a) and he even uses terms like 'information warfare.'[69] According to Volkov the goal of the Russian government is to deliberately spread disinformation and discredit the Russian opposition in Europe by using various low-tech (e.g., pranksters) and hi-tech means. The upshot of this incident is the increased level of awareness of the new techniques being deployed by the Kremlin, according to Volkov (NOS Nieuws 2021a). In sum, the considered judgment for Volkov is that he implicitly rejects the use of these techniques but he is not naïve and knows that these tools and techniques are being deployed by state actors to pursue their own interests.

Alexei Navalny

Since Navalny is unable to speak for himself, I will assume that his considered judgments in this case match Volkov's considered judgments. In addition, I am assuming that Navalny will claim this event to be one of the many tools and techniques that Russia has deployed and will deploy to discredit himself and his organization.

[68] The interview with Volkov took place *before* the Russian pranksters disclosed their video on Youtube on 27 May 2021.

[69] Original quote is in Dutch and has been translated to English by the author.

Dutch Parliament Foreign Affairs committee

The considered judgments for the Dutch Parliament Foreign Affairs committee are applicable to the whole Dutch Parliament, as chairperson Vera Bergkamp alluded to in an interview that this incident could have been prevented; the event has led to new and adapted operating procedures for members of parliament (NOS Nieuws 2021b; Brouwers & Verhagen 2021). According to Kati Piri, one of the members of the committee, this incident has been a wakeup call for them in two ways; initially they thought they were fooled by a deepfake so this made them aware what the future has in store and second, they have become aware that they need to be vigilant in whom to trust, and cannot simply assume by default, that the party they will meet are who they say they are (Van Assen 2021). The considered judgments regarding the ethical issue is derived from the consulted sources as these are implicitly present in these sources; what was written in the sources was primarily based on the parliament's introspection rather than on condemning the behavior of the pranksters. The judgment is that it is not illegal to impersonate somebody if it is satire, however the Dutch parliament has become aware of the potential dangers of deepfake technology and is operating with a more suspicious, vigilant mindset going forward.

Russian comedians

The Russian pranksters Vovan and Lexus, or Vladimir Kuznetsov and Alexei Stolyarov which are their real names, admitted in an interview on 30 April 2021 that they did the pranks with various national governments (Vincent 2021). They have a notorious reputation for having pranked dozens of government officials and celebrities around the globe (Shevchenko 2018) and their aim is to "prank high officials and celebrities and to make a lot of fun and publish it to social media" (Vincent 2021). Despite having funny intentions, they have admitted they would never do anything that would harm Putin or Russia (Walker 2016). European critics have pointed out these pranksters often combine provocative questions and out-of-context edited content so the pranked subject will be put in a bad light (Shevchenko 2018). Their considered judgment regarding the ethical issue is that they think it is allowed to impersonate somebody else for the sole purpose of humor and satire.

Russian government

The Russian government is a stakeholder who is involved indirectly, and their considered judgments are implicitly present in the sources (e.g., Walker 2016; Brouwers 2021). Based on these sources I argue that the considered judgments for the ethical issue will be that it is allowed for a state actor to use any means available in the playbook, like deepfakes or cheapfakes, in order to pursue the state's goals. The goals and intentions of Russia are being interpreted by European authorities to deliberately spread disinformation in order to discredit opposition (Brouwers 2021).

Political influencers

There have been numerous, what I call, political influencers responded to this case, and they came up with their judgments. I have selected Marietje Schaake's response as the input for the considered judgments because her response is documented by herself in a column (Schaake 2021) and is what I consider to be a good representation of the public opinion of these influencers. She wrote one week after the conversation with the fake-Volkov took place but *before* the Russian pranksters posted their video on Youtube. Her main argument is that the Dutch Parliament should prohibit the use of deepfakes in election campaigns in order to prevent the strategic spreading of disinformation, uncertainty and doubt. Schaake's considered judgment, which I recasted to the situation *after* the posting of the prankster-video, is that impersonating of a political person can be allowed but not in all situations. It would be allowed for satirical purposes, but it should not be allowed during political campaigns.

6.4 Moral principles and background theories

In the first part of this section the moral principles that govern the considered judgments will be outlined, followed by an account for the relevant background theories and in the last part the principles/ background theories will be connected to the considered judgments.

Principles

For each principle, listed in table 3 below, a brief description will be provided.

Principle	**Description**
Freedom of speech	One of the core freedom-principles on which the EU has been built (EU FRA 2021). In the Netherlands this right is anchored in article 7 of the constitution (Asscher 2002) and entails that anyone has the freedom to express herself publicly without being censored by the government.
Honesty	A property of human behavior that undergirds the hermeneutics of faith in our society and is part of our social contract (see e.g., Cragg 2000). The default position in our society is based on honesty and trust, in other words, by default, any human engagement is based on honesty and e.g., one can trust the person they meet is the person they say they are (Hancock & Bailensen 2021).

Informed consent	A key principle in data privacy (e.g., Nissenbaum 2011) and bioethics (e.g., Beauchamp 2016) that governs the autonomy of a person so her personal data can only be used *after* explicit consent.
Right to your own face.	A person's face (and voice) are unalienably related to a person's social identity (De Ruiter 2021: 3-4) and therefore cannot be simply copied, used or impersonated.
Schadenfreude	Deliberately making a fool of somebody is justified when it is used as satire. The underlying justification is that this can be "an instrument for social or moral reform" (Bredvold 1940: 256).
Transparency	Clarity of what steps are taken and why in a (political) process; this is required for a democracy to flourish. To fight corruption minority groups often require a higher level of transparency. Also, transparency can increase the level of trust in a democracy (e.g., Schaake 2021).
Sovereignty	Having "supreme control within a territory" (Philpott 2020) and denotes that an entity or person has full political control and is not accountable to other entities like e.g., countries.

Table 3 Moral principles for case study 1

Background theories

The geopolitical situation between the Netherlands (which is a proxy for the EU) and Russia is the primary background theory for this case study. To make a proper ethical assessment, it is important to understand how the geopolitical developments of the last ten years have shaped the current landscape and explains some of the underlying tensions for this case study. The Russian invasion of Crimea and the downing of flight MH-17 have been the two major recent events that have influenced and deteriorated the relations between Russia and the Netherlands. In addition, other incidents like the poisoning of Skripal, OPCW hack, potential election interference and spreading of disinformation keep disturbing the bilateral relationship (Van der Togt 2020). In 2019 the Dutch government has created a Russia-strategy which is based on a "policy of pressure and dialogue, while providing some more options for intensified dialogue and searching for selective cooperation in areas of joint interest" (Van der Togt 2020: 39).[70] It is against this backdrop how the Navalny-case has been interpreted by the Dutch parliament and led to the meeting with Volkov. This background theory

[70] The Dutch Russia-strategy is based on the 2016 EU policy document "Five principles for relations with Russia" (Van der Togt 2020: 36).

exposes the importance of principles like sovereignty and autonomy and why these are extra important in this case. Things would have been perceived and interpreted differently if the pranksters would have come from a different country, like e.g., Germany.

A second background theory that plays a role in this case study is the preconceived notion of deepfakes and the potential harm they can do. Right after the fake Volkov-interview politicians thought they were fooled by a deepfake (e.g., Brouwers 2021; Roth 2021). Vincent (2021) argues that for many years experts have been warning for the infocalypse and the role deepfakes are going to play in this; this has led to fear of deepfakes and this incident is no exception. The major impact for politics has yet to come and Vincent (2021) claims that politicians may blame deepfakes for self-serving reasons, since being fooled by a prankster is more embarrassing than being fooled by technology. The moral implication of this background theory is that it puts the use of deepfakes in perspective and that cheapfakes, like in this case study, can invoke as much harm. The discussions should not be led by fear of technology but should focus more on dealing with truth in a society where truth becomes more and more an ambiguous social construct (Harambam 2021).

Connecting principles and judgments

The judgments-principles connections are listed below in table 4. The principles are grouped by stakeholder and in the third column a brief explanation is added.

Stakeholder	Principles	Description
Volkov	Right to your own face Informed consent Honesty	He has been impersonated multiple times without providing his explicit permission to use his persona. This has resulted in reputational harm.
Navalny	Freedom of speech Transparency	He pursues a higher transparency but is unable to speak for himself as his freedom of speech has been smothered by the Russian government.
Dutch Parliament Foreign Affairs committee	Honesty Freedom of speech Sovereignty	The origin of the meeting with Volkov was indirectly, to support Navalny's freedom of speech. The Volkov-incident has been a wakeup call for them and has been interpreted as an infringement on the basic honesty principle. If it is true that Russia is controlling the pranksters,

		then it is also an attempt to destabilize the country's sovereignty.
Russian comedians	Schadenfreude Freedom of speech	They claim to have the right to entertain and pull pranks (Vincent 2021) which is indirectly based on the freedom to express themselves as comedians.
Russian government	Sovereignty	Russia doesn't want other, especially Western, countries to interfere in their domestic politics.
Political influencers	Freedom of speech Transparency	Marietje Schaake wants to ensure democracy is as transparent as possible to safeguard the trustworthiness of politics. They scrutinize every development that tampers with transparency. They often need to balance this against the freedom of speech principle as the latter is one of the most important cornerstones of our democracy and censoring is unwarranted.

Table 4 Connecting considered judgments and principles for case study 1.

6.5 Reflection

In this section I will provide an account of the reflective equilibrium process that I conducted for this case study. My personal considered judgement regarding the ethical issue is that it is morally impermissible to impersonate somebody especially in high stake engagements like politics. I consider this to be a form of testimonial injustice and illocutionary wronging, even if it turns out to be satire. The main principle that governs my judgment is honesty which is grounded in the biblical Ten Commandments which are an important moral compass to me. I think satire and humor can and must play an important role in our society and that to comedians, freedom of speech is an important *prima facie* right that applies unless it is undermined by an overriding right or principle (Reisner 2013; Ross & Stratton-Lake 2002). For me, after all things considered, for this case study honesty outweighs freedom of speech manifested here as satire, as a moral principle. The main reason for this is that the appearance of the Russian comedians is not recognizable as satire at all and, especially given the sensitive geopolitical backdrop against which this case study takes place, could have resulted in much more harm.

In this case study the conflicting principles are schadenfreude/ freedom of speech versus informed consent/ right to your own face at the personal level of Volkov/ Navalny and sovereignty versus geopolitical power at a country level. The first ethical consideration is if schadenfreude or satire should be banned in the political arena. This to me is not an appealing moral situation since this could easily end up in censorship. In addition, 'politics' is such an ambiguous and value-laden term that it is nearly impossible to come up with a clear-cut definition that determines what is and what is not politics. The opposite of this moral consideration is a situation where 'anything goes' and all kinds and sorts of satire should be morally permissible because it is governed by the freedom of speech. This also would be problematic especially when satire violates other principles like discrimination. The use of satire in politics in the Netherlands is morally permissible under conditions that are governed by the law (e.g., article 1 of the Dutch constitution states that discrimination based on race, sex, religion, belief, and political opinion are prohibited) and existing morality. The latter is constructed and constantly revised by ongoing public debates. The distinction between reality and satire is often a gray area and will become more difficult to grasp, hence that platforms like Facebook are adding labels to social media posts that are designated as satire. This is problematic because who determines what satire is, can these censors or algorithms be trusted and wouldn't this stifle freedom of speech (Hilary & Dumebi 2021: 503). Williams (2014) argues this would lead to maximum gullibility and the whole point of using satire is to keep people sharp and questioning and to take a critical look at the world. In short, I believe satire is an important feature of the Dutch democracy and that the existing laws and morality are sufficiently governing a proper use.

The next ethical issue is whether a person can be impersonated without having provided informed consent. First, I will evaluate this question in general and secondly if the moral stance changes if it is satire. In the Netherlands It is generally considered morally impermissible, and in some cases illegal, to impersonate a person without that person's explicit consent.[71] If the impersonation has a satirical goal, then it can be morally permissible under certain conditions. A good example of this would be a comedian who plays a political character in a TV-show. Since it is known to everybody upfront the TV-show is satirical, then the use of satirical impersonation is morally permissible. Sometimes the satirical disclosure happens after somebody engages an impersonated person which is morally permissible, in my opinion, if nobody gets

[71] Article 231b of the Dutch Criminal Code (Wetboek van Strafrecht) states that it is illegal to use somebody else's identity if this conceals your personal identity and harms the person whose identity has been used. Official law (in Dutch) can be found here: https://wetten.overheid.nl/jci1.3:c:BWBR0001854&boek=Tweede&titeldeel=XII&artikel=231b&z=2017-03-01&g=2017-03-01.

harmed. The Volkov-impersonation by the Russian comedians takes place in a moral grey area. On the one hand during the course of the conversation it becomes clear that the politicians may have been pranked (Van Assen 2021) and nobody has been harmed directly; one could argue this should be allowed in an open democracy that values freedom of speech. On the other hand, given the aforementioned background theories, one could argue that the real Volkov and indirectly Navalny, have incurred reputational and testimonial harm and this prank could have caused more political damage. After evaluating both *pro tanto* arguments I assess that the moral principle of honesty and sovereignty prevails the schadenfreude/ freedom of speech principles and that the satirical impersonation is morally impermissible, and I don't have to revise my considered judgment and principle for this case study.

The ethical issue at macro-level is about to what extent a country may geopolitically influence other countries without compromising the other country's sovereignty. Russia has a long history and track record in using disinformation as a tool to influence other countries (see e.g., Schick 2020: ch. 2; Van der Togt 2020; Brooks 2021). Politically influencing other countries is absolutely morally permissible and is something every country does, however it is only permissible if it is done through generally accepted means like diplomacy or via supranational organizations like the United Nations. It is assumed by European authorities and political analysts that Russia has deployed these comedians (Shevchenko 2018; Van der Togt 2020) as a media tool. If this is true, as Russia would never confess this of course, than this is morally impermissible conduct that infringes Dutch sovereignty. In sum, the geopolitical backdrop explains the sensitivities that are associated to this case study and influence the moral decisions. On a macro-level using deceiving methods is generally considered impermissible.

6.6 Conclusion and lessons learned

After going through the reflective equilibrium process for this case study I have reviewed and assessed the various conflicting principles as a thinker. This provides a good overview of the underlying principles and their mutual which will be helpful for future ethical issues and provides the foundation for the lessons learned, which are summarized in this section. The ethical consideration exposed that satire and freedom of speech need to be balanced against privacy and bodily integrity (right to your own face). In this case study my conclusion as a thinker is that it is morally impermissible to impersonate Volkov, not just because he had not consented to using his face, but also that he was harmed as a speaker and Navalny is done testimonial injustice. The first lesson learned from this case study is that satirical impersonation should always be contextualized in terms of geopolitical, cultural, and testimonial background when

it is being morally assessed. It is important to keep in mind that satire plays an important role in a democracy and in journalism as it “keeps them on their toes” (Van Doorn et al. 2021: 69). The second lesson learned is that events like that happened in this case study will increase the awareness in society that reality altering technology will play a more important role in the (near) future. As a society we need not only to be aware of this but events like this help to start and progress the public discussion about what is morally acceptable and how we can prepare to be more vigilant without becoming cynical.

7. Case study 2. Seeing is (not) believing

7.1 Ethical issue

A short summary of this case: the Belgian environmental pressure group *Extinction Rebellion* (XR) Belgium has posted a deepfake video of the Belgian Prime Minister Sophie Wilmès in which she states that the COVID-19 pandemic is clearly caused by climate change. At the end of the video, it is clearly stated the video is a deepfake, however the content delivered by Wilmès is based on true scientific facts which are published on XR's website along with a verbatim copy of the speech (Extinction Rebellion 2020a). It is clear that XR has used this deepfake as a propaganda tool to pursue their political goals, which is to raise awareness for climate change. What makes this case study also interesting in the context of this paper, is that early 2020, the British counter-terrorism police had identified XR as an organization holding extremist views (Cassam 2021: 1). Extinction Rebellion is an organization that uses non-violent civil disobedience activism and organizes events that attract many school-age children and adults. The concern for climate change is not extreme in itself but it "may encourage vulnerable people to perform acts of violence, or commit such acts themselves," according to the British police (Dodd & Grierson 2020). A year later British politicians and policy makers have admitted it was wrong to label XR as extreme (Wall 2021), despite they still consider XR's events disturbing and annoying.

XR is not a political group holding extreme beliefs but considers itself a group that is voicing an opinion that is not heard enough in mainstream politics, so for this thesis I consider their voice as a minority voice. However, it is a fine line between activistic civil disobedience and considered to be a group holding extreme views (according to the authorities), which make the lessons learned in this case study also applicable to other situations and contexts of extremism and conspiracy theories (see e.g., Katsafanas 2020; Szanto 2020; McCauley & Moskalenko 2017). The ethical issue underpinning the reflective equilibrium process for this case is to morally assess if political actors are justified to use deepfakes, in order to pursue their political goals, and who is responsible for the harm being done.

7.2 Stakeholders and interests

Table 5 below contains an overview of the relevant stakeholders, their role, and their interests for case study 2.

Stakeholder	Role and Interests
XR Belgium	Belgian branch of the worldwide activistic climate movement whose main role is to raise attention for the impact of climate change and aggressively transform the society to become climate friendly.[72]
Sophie Wilmès	Belgian Prime-Minister from 2019 to 2020 and at the time the deepfake was published, was deemed responsible for Belgian climate politics. She is not affiliated to XR Belgium and is virtually, meaning not physically and personally, involved in this case because the speech in the deepfake video is falsely attributed to her (Ray 2021: 986).
Creator of the video	The XR Belgium volunteer who used his/ her technical skills demonstrates that it is relatively easy to create and deploy deepfakes of this level.[73]
Belgian population	The video is targeted towards them, and their role is to assess and judge the political message on its merits. This message may inform or change their voting behavior.
Belgian government	XR Belgium is addressing them. Their role is to govern Belgium and define climate politics. They may or may not respond to messages like this.
Technology providers	Collection of organizations who provide technical services that enable to create, produce, and distribute this deepfake video.[74] They have no direct interest in this case other than that their tools and services enable the creation of this deepfake.
Population of other countries	People in other countries watch this video and this may indirectly influence them.
Future generations	XR is appealing to the climate crisis that will have moral implications for the people who inhabit the earth in the (near) future.

[72] The three main goals of XR Belgium, or demands as they call it, are listed on their website: "WE DEMAND [1] that the Government declares a **Climate and Ecological Emergency**, recognising the need for rapid transformation of our economic system. [2] that the Government enacts a comprehensive, legally-binding **National Emergency Plan** which phases out the extraction and import of fossil fuels by 2025, and prioritises restoration of biodiversity and the preservation of our natural environment. [3] a **Citizens' Assembly**, equipping our regions and communities with the resources and the authority to ensure a managed transition to an equitable post-growth society" (Extinction Rebellion 2021a).

[73] It must be noted that the level of technical expertise of this volunteer is not published. For this thesis I take the assumption that if a volunteer can create this level of videos, it is relatively easy to do.

[74] Think of organizations that host XR's website and software providers who create the tools that are used to create this deepfake video.

Nature	This is a non-human stakeholder representing all what is potentially at stake because of the climate crisis. In a modern conception this entails human, non-human and artefactual entities (Gellers 2020: 120) and is also commonly referred to as environment.

Table 5 Overview of stakeholders, roles and interests for case 2.

7.3 Considered judgments

All relevant considered judgments can be found below, grouped by stakeholder.

XR Belgium

XR Belgium's main interest for deploying this deepfake is to gain visibility for their political agenda and they needed "a strong way to make the government and the population react" (Hulobowicz 2020) on their message. XR's considered judgment is that the looming climate crisis will be much more devastating than the current COVID-19 crisis, that it is justified to use a deepfake to pursue their political goals. According to a XR spokesperson all necessary precautions were taken: "we did a lot of pedagogy and made it clear that it was fiction" (Hulobowicz 2020).[75] By putting up these measures XR Belgium is confident that what they have done is morally responsible and they even take responsibility for the confusion this video caused, or to quote the XR spokesperson again, "If there is any confusion, we deplore it and whenever possible we try to rectify the misperception of people by writing to them that it is a fiction" (Hulobowicz 2020).[76]

Sophie Wilmès

As per the content of the video one might think that Sophie Wilmès agrees with XR's political agenda, from which it follows that her considered judgment would be similar to XR's considered judgment. However, Wilmès does not play an *active* role in this case and her response to this deepfake has not been published[77] so I have to make assumptions on what her considered judgment on this case would be. In this deepfake Wilmès has been digitally impersonated and has been victim of what Diakopoulos and Johnson (2020: 11) call *persona plagiarism*, which they define as "an inversion of plagiarism focused on the source rather than the content of a message." Wilmès's persona (her likeness and voice) is being used by XR Belgium who is exercising control

[75] Quote is originally in French and has been translated to English using Google Translate.
[76] See previous footnote.
[77] XR has shared this video with Sophie Wilmès, the Belgian Climate, Health, and Economy ministers and all the other political party leaders in Belgium. According to XR only the Groen Party and MR (the liberal party of Wilmès) has responded, but the responses are not documented (Extinction Rebellion 2020b). It is unknown whether Wilmès has responded personally or in her role as Prime Minister of Belgium.

over this, without recognizing her as the "other" which makes this problematic (De Ruiter 2021: 15). Ray (2021: 986) argues that "to falsely attribute a speech to an elected Prime Minister is of grave concern" and it is also up for debate whether the video is a parody or not. If it is clearly parody (like e.g., the King Willem-Alexander and Queen Maxima movies from Dutch video-artist Sander van de Pavert[78]) then there is no misattribution and it does not do any harm (Diakopoulos & Johnson 2020). Based on this background information I infer that Sophie Wilmès's considered judgment is that it is impermissible to misattribute her persona for XR's political goals, both in her role as prime minister as personally.

Creator of the video

The only known about the creator of the video is that she is an XR volunteer. From this it follows that her considered judgments will overlap with those of XR. In addition, I will infer that she is willing to deploy her technical skills in order to pursue XR's goals. The reason for bringing up the creator as a separate stakeholder in this case is the role she has in terms of accountability. In moral evaluations of deepfakes there is a distributed morality (Floridi 2013) which means that responsibility is often distributed among various stakeholders involved. The creator of persuasive technology like this deepfake also does have a certain responsibility for the consequences (Coeckelberg 2009; Verbeek 2006). In sum, the considered judgment for the creator is the same as that of XR complemented with the willingness to be partly responsible for the consequences of creating a deepfake.

Belgian population

It is nearly impossible to come up with a single considered judgment for the Belgian population because not much has been published about this. For this thesis I will base the considered judgment on the general opinion that is published in Hulobowicz (2020), Galindo (2020) and the responses to other political deepfakes that took place in Belgium (Von der Burchard 2018; Vooruit 2018)[79] and the Netherlands (Mommers & Wijnberg 2021)[80]. Since this is the first time people really get exposed to a political

[78] For examples (in Dutch) see https://luckytv.nl/willy-en-max/.

[79] In 2018 the Flemish Socialist Party *Vooruit* posted a deepfake video of Donald Trump (Vooruit 2018) where Trump urged the Belgian people to follow America's example and leave the Paris climate agreement (e.g., Rini 2020; Giansiracusa 2021; Von der Burchard 2018). The quality of the deepfake video is not very professional and if you watch it, it is obvious that it is not Donald Trump speaking. The Belgian party intended to use this video to provoke the Belgian people to a conversation about climate change (Von der Burchard 2020) and did not intend to do harm or fool anyone, however, there have been commentators and people who were confused by this video and took the video for real (Giansiracusa 2021: 52).

[80] On 28 October 2021 the Dutch online magazine *De Correspondent* (Mommers & Wijnberg 2021) published a deepfake video of prime minister Mark Rutte in which he voiced the magazine's opinion about how the climate crisis should be tackled in their opinion. They accompanied the video with the hashtag #climateleadership and made it clear at the end of the video that it was a deepfake. In addition, the magazine published an extensive justification on the rationale of their decision to make a deepfake. The video went viral

deepfake, it generates a wide variety of responses from people supporting this video and seeing deepfakes as an interesting new media tool to people who think deepfakes are so bad that they should be banned completely. Despite that the video is clearly marked as fake, a minority is still confused and believes the video is real (Hulobowicz 2020). The considered judgement for the Belgian population is that the majority recognizes the video being fake but rejects the use of deepfakes for politics. From a responsibility perspective it is important to explore if viewers who think the video is real are justified to believe this and XR can be held responsible for this, or that their gullibility is an epistemic vice, and it is their own responsibility (Cassam 2019b).

Belgian government

The Belgian government is represented in this video by Sophie Wilmès, and at the same time they are the target audience for the message that XR wants to convey. Like Wilmès, the Belgian government doesn't play an *active* role in this case study and there haven't been any official responses to this deepfake published. It is still important to have the government present as a separate stakeholder in this case study because they hold the political power in Belgium, and it is very likely deepfakes will play a more important role in the near future (Schick 2020). The considered judgment for the Belgian government in this case study would be that it is impermissible to impersonate official government representatives and let them speak an alleged official government statement when the goal is to promote the official political agenda of XR Belgium. Despite the video being clearly labelled as fake, it still may cause confusion and it is not a clear example of parody. I have inferred this considered judgment on how a government would have responded in similar cases, like e.g., case study 1 in this thesis, where using deceiving and manipulating technology is being rejected by the Dutch government.

Technology providers

Technology providers are the hardware and software vendors that enable the creation, editing and distribution of this deepfake. This would entail e.g., the manufacturers of the deepfake software and the provider that hosts XR's website. These parties don't play an *active* role in this case study but do have, like the creator of the deepfake, a certain level of responsibility. Creators of synthetic media/ deepfake technology have a higher degree of responsibility than e.g., the party hosting the website as the adage goes 'with great power comes great responsibility.' Many creators are aware of this and take their responsibility and ethics very seriously. A good example of this is the *Ethics in Synthetic Media* guide that has been created by the

on social media and the general public opinion varied from a 'cool video' to 'deepfakes are scary and should be forbidden.'

AITHOS Coalition (Aithos 2019), which is a coalition of synthetic software companies. They claim in this guide "As technologists, we recognize our roles in creating new world possibilities and with those exciting advances, we must also recognize the responsibilities that come with innovation" (Aithos 2019: 2). The guide provides ethical recommendations for each software company to consider; in general, all synthetic software companies have an extensive ethical section on their website outlining the ethical terms and conditions how their software can be used.[81] They clearly reject unethical application of their software and will prevent unethical usage as much as possible. It must be noted that a lot of deepfake/ synthetic media software is created and distributed based on an *open source* model, which makes it possible that it can be downloaded and used by everyone in any way they like.[82] The XR deepfake video is most likely created using open source software as, based on the Aithos guidelines, this would have been designated as an unethical application because it uses software to falsely let a public person say something (Aithos 2019: 3). There can be two considered judgments inferred, the first one is for the technology providers for this case study and the second one is specifically for synthetic media software companies.[83] The first considered judgment is that it is permissible to create this XR deepfake since the content is not illegitimate and the deepfake is clearly labelled as fake. The chief responsibility for the content lies with XR. The second considered judgment is that it is impermissible to create this deepfake video because Sophie Wilmès is being harmed and wronged as a speaker and as a public figure. Even with the video being clearly labelled as fake it is still unethical.

Population of other countries

People in other countries than Belgium have watched (and will watch) this deepfake video and be exposed to the message. It can be safely assumed they there will be a similar set of responses to this video than that happened in Belgium with the difference that foreigners (i.e., people not living in Belgium) will interpret this video based on their own cultural background.

Future generations

In climate ethics future generations are often involved as a stakeholder for ethical deliberations and is often referred to as intergenerational justice or ethics (e.g., Gardiner 2010; Spannring 2021). For this thesis it means that the interest of future

[81] For examples see e.g., the ethics webpages of the following synthetic media software companies: *Synthesia* https://www.synthesia.io/ethics and *ReSpeecher* https://www.respeecher.com/ethics.
[82] Examples of open source deepfake software are *FaceSwap* https://faceswap.dev/ and *DeepFaceLab* https://github.com/iperov/DeepFaceLab.
[83] For simplicity I assume there are two technology providers involved in this case study, being the website hosting provider and the open source deepfake software provider.

generations will play a role in the ethical evaluation and will add additional justification to warrant the use of a deepfake in this case study.

Nature

Since Christopher Stone (1972: 456) in his seminal article suggested to grant (legal) rights to nature as a whole, many ethicists, legal scholars and environmentalists have debated to how and to what extent rights and personhood, both legal and moral, should be extended to nature as a whole, referred to as *ecocentrism*, or to all living things that possess inherent worth, referred to as *biocentrism* (for an in-depth discussion see Gellers 2020: 108-117). Adding nature or other non-human entities to the moral circle[84] is based on the premise that the Cartesian dualistic dichotomy between man and nature is illusory (Gellers 2020: 104). For this thesis I will presuppose that nature as a whole is a non-human stakeholder who is involvement in this case study is indirect. The choice for using nature as a separate stakeholder in this case study is to underscore that the interests of non-human entities are becoming more prevalent in ethical considerations.

7.4 Moral principles and background theories

In table 6 below the moral principles that govern the considered judgments will be outlined, followed by an account for the relevant background theories.[85] In the last part the principles/ background theories will be connected to the considered judgments.

Principle	Description
Ecological justice	Rich industrialized countries must take their moral responsibility in radically solving the climate crisis to phase out "the extraction and import of fossil fuels by 2025, and prioritises restoration of biodiversity and the preservation of our natural environment" (Extinction Rebellion 2021a). This should be shaped by principles of equality, freedom, and human rights for all people across the world (Extinction Rebellion 2021b).
Urgency	The climate change is such a grave danger for our society that utmost speed is required, and the government needs

[84] The moral circle is a representation of all the entities a person or institution cares about when performing moral deliberations. According to Singer (1981) this circle is expanding when one grows up. See for a discussion on expanding the moral circle towards nature versus other humans Rottman et al. (2021).
[85] For the principles in this case study that are the same as in case study 1 the content of the *Description*-column will be similar. For legibility I have chosen to use the full description in this table and not to refer to case study 1.

	to come up with an emergency plan to tackle this (Extinction Rebellion 2021a).
Civil disobedience	Defined by Rawls (1971: 364) as "a public, non-violent, conscientious yet political act contrary to law usually done with the aim of bringing about a change in the law or policies of the government." This leads to conflicting duties; on the one hand compliance with the law and on the other hand, the duty to fight injustice (Rawls 1971: 363). Non-violence leads to a higher acceptance and sympathy from the general public and gains credibility with government (Molinari 2021: 31).
Freedom of speech	One of the core freedom-principles on which the EU has been built (EU FRA 2021) which is anchored in the Belgian constitution (Belgium 2016) and entails that anyone has the freedom to express herself publicly without being censored by the government.
Transparency	Clarity of what steps are taken and why in a (political) process; this is required for a democracy to flourish. To fight corruption minority groups often require a higher level of transparency. Also, transparency can increase the level of trust in a democracy (e.g., Schaake 2021) and help to tell the truth (e.g., Extinction Rebellion 2021a; Mommers & Wijnberg 2021).
Informed consent	A key principle in data privacy (e.g., Nissenbaum 2011) and bioethics (e.g., Beauchamp 2016) that governs the autonomy of a person so her personal data can only be used *after* explicit consent.
Right to your own persona	A person's face and voice are unalienably related to a person's social identity (De Ruiter 2021: 3-4) and therefore cannot be simply copied, used, or impersonated.
Credulity	"— in absence of counter-evidence — we should believe that things are as they seem to be" (Swinburne 2004: 293). This is a virtue, closely related to the trust-default (Hancock & Bailenson 2021: 150) in our society. If the believer can be held responsible for holding false beliefs it would be called gullibility.[86]

[86] Gullibility can be defined as foolishness or naiveté. The opposite of gullibility is trust, which can be defined as "believing others in the absence of clear-cut reasons to disbelieve" (Rotter 1980).

Responsibility	Taking a position in product[87] design as to what potential consequences of using the product there are in terms of potential harms and benefits, and to what extent one is accountable and responsible for this.[88]
Liability	Being held (partly) guilty for the harms a product has caused. Is often codified in legislation.
Sovereignty	Having "supreme control within a territory" (Philpott 2020) and denotes that an entity or person has full political control and is not accountable to other entities like e.g., countries.
Sustainability	The moral principle to keep the environment livable and inheritable for future generations.

Table 6 Moral principles for case study 2.

Background theories

In this case study two background theories, COVID-19 and the preconceived notions regarding deepfakes, are relevant and influence the ethical issue at hand. I will only provide an account for COVID-19 since the latter is already covered in the previous chapter (see §6.4).[89] At the time the deepfake video was created (April 2020) it was in the beginning of the COVID-19 pandemic and, like many other countries, Belgium was in a state of lockdown which was unprecedented. Many academics, political parties, and pressure groups (like XR) were making claims about what was the root cause of this pandemic and how the world should change after this pandemic was over (for examples see e.g., Rivera-Ferre et al. 2021; Alcántara-Ayala et al. 2021; Extinction Rebellion 2020a). According to XR the root cause for the pandemic was clearly rooted in climate change. The COVID-19 pandemic has been an important background theory informing and influencing many moral issues in 2020 and is therefore also relevant for this case study.

Connecting principles and judgments

The judgments-principles connections are listed below in table 7. The principles are grouped by stakeholder and in the third column a brief explanation is added.

[87] Product is conceptualized here in a broad sense and also entails digital products (like software) and services.
[88] For a discussion on product safety, responsibility, and liability, see e.g., Moriarty (2021), Coeckelbergh (2006), and Verbeek (2004).
[89] It goes without saying that I will include the background theory for preconceived notions of deepfakes in the reflective equilibrium process.

Stakeholder	Principles	Description
XR Belgium	Ecological justice Urgency Civil disobedience Freedom of speech Transparency	They want to execute on their mission and gain the highest level of campaign visibility (Hulobowicz 2020) and urge the Belgian government to take steps now and tell the truth about climate change (Extinction Rebellion 2020a).
Sophie Wilmès	Informed consent Right to your own persona	She did not approve in advance for her face and voice to be used by XR. This applies to both her embodied self as well as her role as prime minister.
Creator of the video	Same principles as XR Responsibility	The XR volunteer accepts being responsible for the risks involved when deploying deepfake technology.
Belgian population	Credulity Transparency	Their default hermeneutical mode for interpreting videos is that of believing what they see, especially when the main character in the video is being portrayed in her official role as prime minister.
Belgian government	Credulity Transparency Sovereignty	Belgian government principles overlap that of the Belgian population (see row above). In addition, this deepfake undermines the authority of the Belgian government as this is not an official statement of the government.
Technology providers	Freedom of speech Responsibility Liability	The first considered judgment (provides related to this case study) have limited responsibility and claim their users can use the technology to express themselves and they are themselves responsible for this. The second considered judgment (synthetic media software providers) is driven by ethical business conduct and considers this deepfake unethical, so this will not be made possible on their platforms.
Population of other countries	Credulity	Their default hermeneutical mode for interpreting videos is that of believing

		what they see and applying this to their own cultural context.
Future generations	Ecological justice Urgency Transparency Liability Sustainability	The impact of climate change will be prevalent for this group. This will increase the moral weight for the justification for the methods used in this case study which is morally permissible under some circumstances (Gardiner 2010: 94).
Nature	Ecological justice Urgency Sustainability	If it is true that nature should have rights, then it follows that the interests of nature get involved in the reflective equilibrium process. Its main governing moral principle is sustainability, in other words, that nature can keep running its course and that a moral case also takes the interests of non-human entities into ethical consideration.[90]

Table 7 Connecting considered judgments and principles for case study 2.

7.5 Reflection

In this section I will provide an account of the reflective equilibrium process that I conducted for this case study. My *personal* considered judgment as a thinker is that it is morally impermissible to use deepfakes for impersonating politicians as an activistic method for organizations to pursue their political goals. The reason is that the deepfaked person is being wronged in their speech-act and hasn't consented to using her persona. In my opinion every human being, regardless of what role she has, is entitled to her own persona and controls how this is being used. The purpose of the deepfake video is clearly political and not satire, despite the video is being labelled as a fake video. In this case an unethical means is being used to pursue a political end. The main principle that governs my judgment is honesty which is grounded in the biblical Ten Commandments which are an important moral compass to me. The conflicting principles for this case study are the following: (i) can freedom of speech, non-violent civil disobedience, the urgency of the climate crisis, intergenerational justice and the rights of nature be used as reasons to justify the creation of this

[90] It is still provocative to include non-human entities like nature in ethical and legal issues, however this view is getting more traction. There are many documented and published examples where elements of nature have been granted legal rights (for examples see Gellers 2020: ch. 4; Borràs 2016).

deepfake that wrongs a person's right to their own persona without given prior consent, (ii) can a deepfake be used as an act of non-violent civil disobedience if it causes confusion and wrongs the credulity of the Belgian population,[91] and (iii) to what extent are technology providers responsible for the content (in this case the deepfake) that is created or hosted on their platforms.

This deepfake was created at the beginning of the COVID-19 pandemic which was a time of unprecedented extremes. The world was taking by surprise and governments around the globe had to take extreme measures, like lockdowns and curfews, in order to control this pandemic. Given this backdrop of extreme times it can be assumed that activist groups will run extreme campaigns in order to create attention and raise awareness for their goals. Given this context XR partly justified the creation of this deepfake. According to a XR spokesperson, these times require a "strong way" to make the government and the people respond (Hulobowicz 2020). More and more people are buying into XR's philosophy for non-violent civil disobedience and qualify this as rational instead of radical (e.g., Shah 2021).[92] From XR's viewpoint the sense of urgency and the devastating impact of the climate crisis justifies the creation of this deepfake. This sense of urgency is also informed by the impact the climate crisis has for future generations and that XR feel responsible to represent the interest of nature in moral and political public discussions. After all things considered, XR made the decision that their *pro tanto* right of freedom of speech, their *pro tanto* obligation to save the planet outweighs the *pro tanto* obligation (Reisner 2013) of not harming other people, in other words, the harm incurred by Sophie Wilmès, the Belgian government and population is the price they are willing to pay. A mitigating factor is that XR has done everything in their power to mitigate the risks associated with the release of this deepfake (Hulobowicz 2020). Most harm is done to Sophie Wilmès, both personally and in her role as prime minister. First, her persona is being used without her explicit consent (persona plagiarism). It follows from modern psychological, medical, and ethical literature that the face and voice of a person are unalienably related to the identity of this person which also applies in a digital environment. De Ruiter (2021) cogently argues that a person has a right to digital self-representation, from which it follows that in case of non-consensual deepfakes "others may not manipulate digital data that represent people's image and voice, as markers of the self, in hyper-realistic footage that presents them in ways to which they would object" (2021: 16). Second,

[91] I will also briefly discuss s hypothetical situation in which XR is considered by the authorities as an organization holding extreme beliefs (which is not such an extreme thought as this happened in the UK (Cassam 2021: 1; Dodd & Grierson 2020). Case study 3 will solely focus on a case where a deepfakes is used by an organization holding extreme (conspiracy) beliefs.

[92] A similar rationale was used in the Mark Rutte deepfake, where the creators, who are journalists themselves, used the sense of urgency of the climate crisis to justify the creation of this deepfake (Mommers & Wijnberg 2021).

Wilmès is being harmed as a speaker through illocutionary wronging (Rini & Cohen 2021) as the fabricated speech may not comply with what she would have expressed herself, both personally and in her role as prime minister. Despite Wilmès's public role as prime minister which comes with a certain amount of wronging and satire, I think it is wrong to non-consensually use her persona in any case. In this case the *pro tanto* principles 'informed consent' and 'right to your own persona' after all things considered prevail the principles used by XR. Non-consensual impersonation is morally impermissible because fundamental human rights, which are anchored in the Belgian constitution, are wronged and political pressure groups like XR have other means that are justified to pursue their political goals.

The second conflict of principles is between the principle of using non-violent civil obedience that can cause confusion and the credulity of the Belgian population. XR's intentions were transparent from the start and, according to a spokesperson, they have done everything in their power to make sure the audience knows the video is a deepfake and they expressed their willingness to reach out to people who got confused (Hulobowicz 2020). XR has deliberately chosen for using a deepfake because the technology was convincing and, according to their spokesperson, the society has already crossed deepfakes without their knowledge (Hulobowicz 2020). The problem with this last statement is that the use of deepfakes was not very common in 2020, however, it is true that the Belgian population, as well as the population of other countries, have been exposed to manipulated content like photographs before. The deepfake did cause confusion with a minority of the Belgian population (Hulobowicz 2020) which wronged their credulity. I think that the XR deepfake did not wrong the credulity of the total Belgian population because it was clearly labelled as a deepfake. One could argue that the minority who got confused was gullible because the message was too good to be true and they ignored the clear signs that the deepfake was fake. A side effect of using deepfake technology is that it provokes a public discussion and raises awareness of its potential benefits and harms. I think this is a necessary development in order to prepare for a society where deepfakes will become more prevalent. One of the ways that a society can prepare is through increasing media literacy both in schools and in public information. Media literacy can help to reduce gullibility and enables a society to deal with deepfakes in responsible manner.[93] In sum, XR is justified to use a deepfake as non-violent civil disobedience because they have been fully transparent in doing so and clearly labelled the video as fake. The Belgian population, and the population of other countries, could have known this and

[93] A good example of an organization whose goal is to increase media literacy in a world of deepfakes is human rights organization WITNESS. They organize workshops for journalists, opinion leaders and civil rights activists how to deal with deepfakes and they create many resources that enables the educate the general public (WITNESS 2021).

are not harmed in their credulity. But what if XR would hold extreme beliefs and had malicious intentions. Suppose, *reductio ad absurdum*, that XR Belgium would want the Belgian government to step down and be taken over by representatives of XR and they created a deepfake in which Sophie Wilmès announced to step down and hand over all governmental power to representatives of XR. In this case the Belgian population cannot be blamed of gullibility because they are justified to think the content in the video is right as it is not labelled as a deepfake, and it is not satire. I think in this case the use of a deepfake is morally impermissible because XR is deliberately trying to create confusion and uses illegal and unethical means to pursue their political goals. In short, when extreme beliefs inform malicious intents of using deepfakes this has huge consequences for moral and political deliberations.

The third conflict of principles is about whether technology providers can be held responsible and liable for the deepfakes that are being created and hosted on their platforms. The conflict is between the following principles: the freedom of expression which leads that the creator takes full responsibility for the consequences and the (partial) responsibility of the technology provider for the consequences of the deepfake. For synthetic media software providers this case study would be morally impermissible because the video created is non-consensual. When I interviewed Victor Riparbelli, the CEO and co-founder of Synthesia.io, and asked him about this case study, he pointed out that their website clearly states the following: "We will never re-enact someone without their explicit consent" (Synthesia 2021). I have assumed this XR deepfake has been created using open source software. Usually open source software, just like any software, comes with a license agreement that states the terms and conditions how the software can be used.[94] Usually this license agreement limits the liability for the software provider making the user fully responsible for the content created. For this case study it is fair to say that XR carries the full responsibility for this video and the technology providers[95] are not responsible and liable for the consequences of this deepfake because the deepfake complies with the terms and conditions of the technology providers and the content is not showing anything extraordinary or illegal.[96]

[94] An example of this would be the *GNU General Public License* which is a common open source license framework. By downloading and installing the software one agrees with the terms and conditions that are stated in this agreement. An example of this license agreement can be found on the *ml-deepfake-GAN* website https://github.com/as-ideas/ml-deepfake-GAN/blob/master/LICENSE.

[95] For simplicity I assume there are two technology providers involved in this case study, being the website hosting provider and the open source deepfake software provider.

[96] With large (social) platforms like Twitter and Facebook, there is an ongoing discussion as to what extent they can be held responsible for the content on their platform or not. In the United States the government is considering a revision of Section 230 of the *Communications Decency Act*, which exempts platforms for being responsible for the content that is being published on their platform (see e.g., Ashford 2021; Cheng & Norcross

7.6 Conclusion and lessons learned

After going through the reflective equilibrium process for this case study I have reviewed and assessed the various conflicting principles as a thinker. This provides a good overview of the underlying principles and their mutual which will be helpful for future ethical issues and provides the foundation for the lessons learned, which are summarized in this section. As a reminder, the ethical issue is to morally assess whether political actors are justified to use deepfakes, in order to pursue their political goals, and who is responsible for the harm being done. In general, my conclusion is that *in itself* it is morally permissible for political actors to use deepfake technology to pursue political goals, however for this specific case study I have concluded after all things considered, that non-consensual re-enactment of a person in a deepfake is outweighs XR's *pro tanto* principle of freedom of speech, their appeal to climate crisis urgency, intergenerational justice and sustainability. The first lesson learned is that consented use of personas should have significant weight in future ethical deepfake evaluations, especially when a political organization has alternative means at their disposal to pursue their political goals. Second lesson learned is that events like this case study help to push the public discussion about deepfakes forward. It helps to conceptualize what deepfakes are and creates precedents about what we, as a society, think is morally acceptable. An interesting observation was made by a commentator in the Dutch newspaper *Reformatorisch Dagblad* on the Mark Rutte deepfake case where he thought the use of deepfake technology was justified because the urgency of the climate crisis (De Jong 2021). To me this is a sign that our society is morally exploring the boundaries of the upcoming deepfake era. The third lesson learned is that ignorance is no longer an excuse for deepfake gullibility under the condition the deepfake is clearly labelled and disclosed as a deepfake. However, this does not mean that everyone in our society already knows what deepfakes are, so I would recommend policy makers and educators to sufficiently pay attention to this phenomenon in the context of media literacy. The fourth and last lesson learned from this case study is that moral responsibility for the consequences of using deepfakes is a complex and distributed problem which is context dependent. For this case study it is clear that XR has, and is willing to take, the full moral responsibility for using deepfakes. For any other case where, moral responsibility might be less clear, it is recommended to discuss and review this in the full context of all stakeholders. With the Mark Rutte deepfake case some people responded on social media that using

2021). In the EU has proposed the *Digital Services Act* to lay out the responsibilities, rights and duties of the digital platforms and users (see e.g., Savin 2021).

deepfake technology is like playing with fire and should be completely forbidden.[97] However, this is not a very realistic option since it is very hard to forbid technology like this and there is a chance one throws the baby out with the bathwater. Responses like this have led me to the following recommendation that it is important to educate people on deepfakes and its consequences in order to have a public discussion on what is morally accepted in our society.

[97] An example can be found in this Twitter thread (in Dutch) where it is claimed that the use of deepfakes is absolutely morally impermissible because of its undermining effects on our democracy and society. https://twitter.com/koertvb/status/1453787222323470353.

8. Case study 3. Will fake be the new real?

8.1 Ethical issue

A short summary of this case: the organization NVR has created and distributed a deepfake in which the prime minister of the Netherlands (Mark Rutte) and a CEO of one of the leading COVID-19 vaccine manufacturers (Maurits Majoor) have a conversation that NVR should be declared an illegal organization, its leader Bill d'Angelo should be arrested and incarcerated and anyone affiliated with NVR should be treated and scrutinized as a potential terrorist. Rutte and Majoor claim the video is a deepfake (liar's dividend), however the Dutch people, who have gotten accustomed to detection algorithms for truth claims, have indicated that the video might be true, and the video led to a decrease in trust of the government and the repeat vaccine. The ethical issue underpinning this case study is to assess the moral implications of a deepfake that is been created and distributed by an organization holding extreme beliefs and conspiracy theories in a world where the population is mainly trusting in technology to assess the truth. The limitation of this case study is that I look at the year 2031 through the lens of the state of affairs in 2021, but since ethics is based on a constant dialogue in society it can be safely assumed that in real life ethics will evolve with the developments in society and the moral principles that are at stake in this case study can be helpful in informing this ethical conversation.

8.2 Stakeholders and interests

Table 8 below contains an overview of the relevant stakeholders, their role, and their interests for case study 3.

Stakeholder	**Role and Interests**
Bill d'Angelo	Leader and co-founder of NVR. His main objective is to let the Dutch population believe the EU and Big Pharma are part of a big conspiracy where they want to control the EU population.
NVR	Activistic pressure group that is firmly believing that the COVID-19 pandemic was a catalyst for the incumbent Big Pharma conspiracy theory (Blaskiewicz 2013)[98] in the Dutch

[98] Blaskiewicz (2013: 259) defines *Big Pharma conspiracy theory* as "the conspiracy theory that pharmaceutical companies, regulators, politicians, and others are secretly working in consort against the public interest." There are many similarities with existing conspiracy theories (e.g., the majority of people are ignorant; a small number of people are working in secret to undermine public good; lack of evidence for conspiracy) but also some special properties like the *cui bono*-claim (who benefits) that people are deliberately held on more expensive and not very effective medication and the general suspicion against vaccines since they are developed and marketed by pharmaceutical companies whose goal is to maximize shareholder value (Blaskiewicz 2013: 260). It should also be noted that Blaskiewicz's definition of Big Pharma is broader than my

	society. NVR claims that EU and Big Pharma are conspiring together and are using repeat vaccines as a means to control the European population and make a lot of money. NVR thinks the vast majority of the EU population is ignorant for this and they want to fight this by all means possible. The Dutch National Coordinator for Security and Counterterrorism has declared NVR as an organization holding extreme views and its members are prone to use any means that may undermine authority or democracy.
Mark Rutte/ Dutch government.[99]	He is prime minister of the Netherlands and in this case study he is representative for the Dutch government. His government is still dealing with the aftermath of the COVID-19 pandemic and its official policy is that an annual repeat vaccination is the only way out of this pandemic. The repeat vaccination is not mandatory but in practice one needs to be vaccinated to partake in social life.
Maurits Majoor/ Big Pharma.[100]	He is the CEO of the Dutch branch of one the leading multinational pharmaceutical companies in the world. His company is headquartered in the UK and the Dutch market is considered to be an important test market for trying out new innovations.
Deepfake detection providers	A public-private partnership between the EU, social media platforms and deepfake detection providers. This partnership has resulted in a deepfake detection algorithm that moderates every video that is being uploaded on the Internet in Europe. The output of the algorithm is a classification whether a video is genuine or deepfake. If the video is classified as deepfake a mandatory deepfake label will be added to the video so it is clear for the viewers that it is a deepfake. A governance board is overseeing this partnership to ensure that the algorithms are attuned to European values. It is mandatory for every internet and software provider in the EU to use this algorithm and this has led to a big reduction of the amount of deepfakes on

definition in this thesis; on his account Big Pharma encompasses not only pharmaceutical companies but also government, NGO's, regulators etc.

[99] In this case study Mark Rutte is representative for the whole Dutch government. He is being harmed as a person because their persona has been used non-consensually. Since this is similar to case study 2 I will not repeat it in this case study. It goes without saying these harms wilbe part of the reflective equilibrium process.

[100] In this case study Maurits Majoor is representative for Big Pharma. See previous footnote about non-consensual use of his persona.

	the internet and an increased trust in the veracity of the videos that are qualified as genuine.[101]
Dutch population	They have learned to trust the mandatory deepfake detection algorithms.

Table 8 List of stakeholders, their roles and interests for case study 3.

8.3 Considered judgments

All relevant considered judgments can be found below, grouped by stakeholder. Since this case study is based on a thought experiment it follows that all considered judgments are inferred from existing literature on COVID-19 conspiracy theories, deepfakes and the interviews I held with the conspiracy theory and deepfake experts.

Bill d'Angelo

Bill d'Angelo's main considered judgment in this case study is that he thinks it is morally permissible to use this deepfake. On his account the EU, Dutch government and Big Pharma are deliberately deceiving the EU population and therefore it is justified to use every non-violent means possible to disclose their true intentions. D'Angelo will of course deny any involvement with this video, and he will capitalize on the fact this video is real and this supports NVR's claim of Big Pharma conspiracy.

NVR

NVR is a European movement that started in the Netherlands and has spread its activities over Europe. Its headquarters is still based in the Netherlands and most of their successes have been achieved there. They have a loyal group of volunteers across all strata of the European population that help them out to execute on their mission. NVR's considered judgment is similar to that of d'Angelo.

Mark Rutte/ Dutch government

Mark Rutte (and the Dutch government's) considered judgment is based on the assumption that NVR has created the deepfake, despite their denial. Rutte is invoking the liar's dividend by claiming this video is a deepfake. On his account the use of deepfakes is morally impermissible because NVR has used unethical means to pursue their political goals. NVR has deliberately created disinformation by creating a deepfake that is not labelled as a deepfake, is making non-consensual use of Rutte's

[101] This presupposition is highly unlikely to take place in real life. Several deepfake experts have told me in interviews that creating a deepfake detection algorithm with this epistemic status is almost impossible to create because of the lack of context the algorithms have. However, for this assertions also applies that they look at the future through the lens of today.

persona, and wronging him in his speech-act and is undermining the trust in the Dutch society and government repeat vaccination policy.

Maurits Majoor/ Big Pharma

Majoor's considered judgment is based on Mark Rutte's assumption that the video is a deepfake. He considers this deepfake morally impermissible because of the same reasons as Rutte and this deepfake causes reputational harm to organizations that created COVID-19 vaccines.

Deepfake detection providers

The deepfake detection providers considered judgment is that it is morally impermissible to create this deepfake. They assume it is a deepfake and that their algorithm has misclassified it as a genuine video. They are looking for any additional evidence that helps them prove this video turns out to be a deepfake, so they can use this to improve their algorithm.

Dutch population

In general, the considered judgment for the Dutch population is that it is morally impermissible to create deepfakes that create disinformation and cause confusion. However, the video causes two different responses among the Dutch people. The first response is based on the people who believe the video is genuine and true and their belief is justified by the deepfake detection algorithm. In addition to NRV adherents who see their beliefs in their echo chamber (Nguyen 2020) confirmed by this video, there is a group who does not support NVR but may find this video compelling evidence for government corruption and may diminish their trust in the government and big pharma. This non-doxastic response endorses the content of the video without buying into NVR's conspiracy theories (Ichino & Räikkä 2020). The second group considers the content of the video 'out of touch' with reality and thinks this is most likely a deepfake.

8.4 Moral principles and background theories

In table 9 below the moral principles[102] that govern the considered judgments will be outlined, followed by an account for the relevant background theories. In the last part the principles/ background theories will be connected to the considered judgments.

[102] For the principles in this case study that are the same as in case study 1 or 2, the content of the *Description*-column will be similar. For legibility I have chosen to use the full description in this table and not to refer to case study 1 or 2.

Principle	Description
Freedom of speech	One of the core freedom-principles on which the EU has been built (EU FRA 2021). In the Netherlands this right is anchored in article 7 of the constitution (Asscher 2002) and entails that anyone has the freedom to express herself publicly without being censored by the government.
Individual autonomy/ bodily integrity	Defined as "the capacity to be one's own person, to live one's life according to reasons and motives that are taken as one's own and not the product of manipulative or distorting external forces, to be in this way independent" (Christman 2020). For this case study this principle can be narrowed down to the notion of bodily integrity. Bodily integrity is violated when the government imposes mandatory vaccination on their population since they non-consensually sidestep "persons' preferences about the handling of their bodies" (Allen 2021).
Transparency	Clarity of what steps are taken and why in a (political) process; this is required for a democracy to flourish. To fight corruption minority groups often require a higher level of transparency. Also, transparency can increase the level of trust in a democracy (e.g., Schaake 2021) and help to tell the truth (e.g., Mommers & Wijnberg 2021).
Civil disobedience	Defined by Rawls (1971: 364) as "a public, non-violent, conscientious yet political act contrary to law usually done with the aim of bringing about a change in the law or policies of the government." This leads to conflicting duties; on the one hand compliance with the law and on the other hand, the duty to fight injustice (Rawls 1971: 363). Non-violence leads to a higher acceptance and sympathy from the general public and gains credibility with government (Molinari 2021: 31).
Informed consent	A key principle in data privacy (e.g., Nissenbaum 2011) and bioethics (e.g., Beauchamp 2016) that governs the autonomy of a person so her personal data can only be used *after* explicit consent.
Right to your own persona	A person's face and voice are unalienably related to a person's social identity (De Ruiter 2021: 3-4) and therefore cannot be simply copied, used or impersonated.

Credulity	"— in absence of counter-evidence — we should believe that things are as they seem to be" (Swinburne 2004: 293). This is a virtue, closely related to the trust-default (Hancock & Bailenson 2021: 150) in our society. If the believer can be held responsible for holding false beliefs it would be called gullibility.
Sovereignty	Having "supreme control within a territory" (Philpott 2020) and denotes that an entity or person has full political control and is not accountable to other entities like e.g., countries.
Accuracy	A metric to measure the performance of algorithms.
Reliable and affordable vaccines for everybody in the world	The mission statement of the alliance of Dutch pharmaceutical companies.[103]

Table 9 Moral principles for case study 3.

Background theories

The main background theory for this case study is the trust in deepfake detection algorithms for detecting the veracity of media. This has moral implications for what is considered to be true or not and this algorithm has proven to be helpful for the Dutch population to navigate in a society that is inundated with synthetic media and deepfakes. To ensure this algorithmic moderation is helpful to the society the EU has mandated that an independent public-private organization creates and deploys an algorithm is being used everybody in the EU.[104] The COVID-19 pandemic and its aftermath has been an important background theory informing and influencing many moral issues in 2031 and is therefore also relevant for this case study.

Connecting principles and judgments

The judgments-principles connections are listed below in table 10. The principles are grouped by stakeholder and in the third column a brief explanation is added.

[103] This fictitious mission statement is inspired by the mission statement of the pharmaceutical company *Pfizer* which in real life is a manufacturer of COVID-19 vaccines. See https://www.pfizer.nl/dossier-corona.

[104] Algorithmic content moderation by third parties is not unproblematic (see e.g., Gorwa et al. (2020) for an in-depth discussion). In this case study I presuppose that there are lessons learned from algorithmic content moderation by commercial internet platforms which led to this mandatory public-private partnership.

Stakeholder	Principles	Description
Bill d'Angelo and NVR[105]	Freedom of speech Individual autonomy/ bodily integrity Transparency Civil disobedience	Their major claim is that the government and big pharma are deliberately misleading the Dutch population and are acting on a secret political agenda of submission and control in which they are violating foundational human rights. They use every (non-violent) means possible to get their message across.
Mark Rutte/ Dutch government	Informed consent Right to your own persona[106] Sovereignty Transparency	Have a legal obligation to govern the country related to healthcare based on a policy of trust and transparency. They will fight anything that is undermining the that jeopardizes the core principles of democracy.
Maurits Majoor/ Big Pharma	Informed consent Right to your own persona[107] Reliable and affordable vaccines for everybody in the world	Their mission is to provide good quality vaccines and to maintain the reputation of Big Pharma.
Deepfake detection providers	Accuracy Transparency	Their goal is to create the best performing algorithms possible, created in a transparent and democratically governed process. Their governance board oversees not only algorithm performance and quality but also that the algorithm is best aligned with EU fundamental human rights (EU FRA 2021).
Dutch population	Credulity Transparency	Their trust-default is based on the outcomes of the detection algorithms.

Table 10 Connecting considered judgments and principles for case study 3.

[105] D'Angelo and NVR are grouped together here because their governing principles are the same.
[106] Both the *Informed consent* and *Right to your own persona* principles are mentioned here for reference purposes and they are part of the reflective equilibrium analysis. I will not go into any details since they are similar how they are applied in case study 2.
[107] Ibidem.

8.5 Reflection

In this section I will provide an account of the reflective equilibrium process that I conducted for this case study. My *personal* considered judgment is that it is impermissible to use a deepfake in this case study because of the following reasons: (i) the video is not marked as a deepfake and is not detected by the deepfake detection algorithms which deliberately causes confusion, (ii) the non-consensual use of the personas of Rutte and Majoor, (iii) civil disobedience like this may be non-violent and legal but it also may incite people to do other acts of civil disobedience that can become violent. The conflicting principles for this case study are the different interpretations of transparency, personal autonomy or bodily integrity versus the epistemic authority of the government and science that mandates the use of repeat vaccines and the epistemic authority of the deepfake detectors versus the credulity of the Dutch population.

Each stakeholder in this case study uses a different interpretation of the moral principle transparency. For NVR transparency means exposing what they think is the concealed truth behind the rationale for repeat vaccination. For the Dutch government and Big Pharma transparency means clearly communicating why repeat vaccinations are necessary to keep the society open and ensure that the Dutch national healthcare system otherwise would collapse. The hermeneutics of transparency is the backdrop for a broad public debate that has been going on for many years and is taking place in newspapers, TV talkshows and on social media. According to NVR this public debate is being hijacked by the Dutch government and Big Pharma to deliberately hide the truth and frame their organization as a potential threat to society since they are holding extreme views. According to NVR it is necessary and justified to use non-conventional means like deepfakes in order to get attention for their message. Their 'guerilla-approach' is a transparency-paradox in itself as it conceals the true intention of the deepfake and the true identity of the deepfake creators while their true intention is to create more transparency. Using technological means for propaganda purposes is nothing new (e.g., Schick 2020, Fallis 2020, Vaccari & Chadwick 2020) and deepfake technology can be considered as a new tool for creating propaganda. Harris (2021: 17) argues that the use of deepfakes refers to more fundamental social epistemic problems:

> The epistemic threat posed by deepfakes is due in part to existing social epistemic crises. Similarly, the other harms threatened by deepfakes have much to do with existing inequalities, prejudices, and the like. While there is a serious threat that deepfakes will reinforce these problems, such applications are not mandated by the nature of deepfake technology.

The existing social epistemic crisis that is underpinning this case study is the diminishing trust in governments and other institutions that hold epistemic authority (see e.g., Baurmann & Cohnitz 2021; Haramabam 2020; Aupers & Harambam 2018). Putting the use of deepfakes in this broader sociological and social epistemological context helps to analyze deepfakes in a social context and not as an isolated technological phenomenon. From an ethical perspective it is morally undesirable to use deepfakes for political purposes, but seeing the use of deepfakes in its sociological context will help policy makers and analysts better understand the phenomenon and not see it in isolation.

The second conflict of principles has a similar line of reasoning as the previous conflict above. NVR claims that the use of non-conventional technology is justified by the far-reaching consequences of the government and Big Pharma's indirect force for getting repeat vaccinations. The underlying social trend that explains this is the epistemic claim of parties like NVR to hold epistemic authority, in other words be an expert, which is based on the abundance of information available on the Internet (Grundmann 2017; Hetmański 2020). This phenomenon logically follows from the lack of epistemic trust in government and institutions as describe above. As argued above the use of deepfakes is morally undesirable but governments should consider this to be a technology that is incumbent in our society and could be weaponized by every activist or fundamentalist group that is withheld from political participation. In this case study NVR is considered by the government to be a group holding extreme views who are "willing to seriously break the law or engage in activities that undermine the democratic legal order" (NCTV 2021). In our technologically mediated society deepfakes should be considered as a potential weapon that can be used by groups holding extreme or fundamental views to pursue their political goals.

The last conflict of principles is to what extent the Dutch population is justified to trust the predicted outcomes of deepfake detection algorithms. In this case study the deepfake detection algorithms are very successful in terms of accuracy and from this it follows that the algorithm's veridicality increases the credulity of the Dutch population. In the 2031 infocalypse one has to trust external tools in order to make a truth claim, however one might question whether the outsourcing of truth assessment to algorithms may lead to gullibility, in other words, in this case of a false positive score is a Dutch person warranted to hold a false belief that the video is true or can it reasonably be expected that one invokes their personal cognitive faculties to assess this deepfake as false. I think it is reasonable to assume that NVR and its adherents are justified to hold their belief in the veracity of this deepfake as it is an extension of their doxastic belief in the conspiracy theories that NVR is disseminating. In addition, there is also a part of the Dutch population that holds a non-doxastic positive attitude

towards this video; on the account of Ichino and Räikkä (2021: 11) there are people who do not believe the underpinning theory of NVR but *hope* this may be true and there are people who endorse this video to simply *communicate* their support for NVR adherents and this video. In my opinion both the doxastic and non-doxastic claims are justified attitudes in this case study. On the one hand it is reasonable and generally accepted to follow the results of the deepfake detection algorithm, in other words this provides both an *epistemic* and *prudential* norm to justify their beliefs (Chignell 2018). On the other hand, it is an example of the *illusory truth effect* (Meckel & Steinacker 2021: 15) where repeated exposure to NPV messaging feeds existing biases, like the confirmation bias (Chesney & Citron 2019). The interesting question is whether the vast majority that doesn't hold these doxastic beliefs or non-doxastic attitudes who believe that the deepfake are justified in their beliefs, in other words, are they culpable for their ignorance. According to Peels (2011: 582) there are two ways that one can be blamed for being or becoming ignorant: (i) the first is based on *akrasia*[108] which might be the case in evidence gathering or working on one's epistemic vices or virtues and (ii) based on "unactivated dispositional beliefs about one's circumstances or the normative status of that action that one should have activated." I think in this case study the epistemic uncertainty that has been caused by the infocalypse could lead to a status of *algorithmic akrasia* where the deepfake detection algorithm provides epistemic certainty despite we know 'in the back of our minds' we may need our existing cognitive faculties. If the algorithm's output is that *p* is true we hold this to be true, unless proven otherwise. Because the accuracy of the algorithm is so good the mental heuristics of the people are starting to trust the algorithm over our cognitive faculties. In sum, I think the majority of the Dutch population can be held responsible for their gullibility based on *algorithmic akrasia*.

8.6 Conclusion and lessons learned

This hypothetical case study is based on a simulacrum (Baudrillard 1994) where reality can be generated by a smartphone. I have used the presupposition this will lead to a truth collapse and that people start to rely more on technology to help them navigate in society. Deepfake detection algorithms will get a high epistemic status in this scenario and generate knowledge about what is real and true. Deepfake technology is not created in a vacuum and one of the lessons learned is to position deepfakes not as an isolated technological category but as a phenomenon that is situated in a social context and is responsive to sociological trends. Leveraging technology for manipulation and propaganda's sake is nothing new and using this

[108] *Akrasia* can be defined as an action or scenario "in which a person does or fails to do something despite occurrently (consciously) believing that doing so is wrong" (Peels 2011: 576).

sociological lens will help policy makers and ethicists to take a broader view and make a better-informed moral assessment. Deepfakes can and will most likely be weaponized by groups holding extreme and fundamentalist views. They will capitalize on the existing sociological trend of diminishing trust in epistemic authorities and leverage this to create dis- and misinformation. This is not new as the European countries have dealt already with Russia as an actor in creating disinformation (see e.g., Schick 2020: ch. 2; Van der Togt 2020). The outcome of the reflective equilibrium for this case study is the awareness that deepfakes are here to stay and going to be more pervasive in the future. Another lesson learned is to not consider deepfakes as an isolated technological category but to see it in broader underlying sociological and epistemological trends. This may cause an infocalypse which on its turn will create a demand for technological solutions, like the deepfake detection algorithms in this case study, to help navigate the epistemic uncertainty. This may sound as an attractive epistemological quick-fix but leveraging technologies to replace fundamental human cognitive skills will become problematic and may cause problems like algorithmic akrasia.[109] Looking at deepfakes in the context of history of manipulation and sociological trends it will help policy makers for technology, extremism and ethics take a broader and more nuanced view on the impact it is expected to have.

[109] Technology is shaping human conditions. A good example is the use of navigation systems like *TomTom* have changed the way people are wired to navigate unknown territory. People have lost their navigational faculties because of this (see e.g., McKinlay 2016).

Conclusion

Are deepfakes going to create social and epistemic turmoil, will our society need algorithms to support our moral assessments, will deepfakes be weaponized by groups holding extreme views or will we, as a society, be able to cope and adapt ourselves to this? The rise of deepfakes brings about intriguing, fundamental questions about what truth, reality and trust mean in a world that is expected to be overwhelmed by a torrent of fabricated and synthetically generated media. The responses in academic literature vary from a dystopian view (e.g., Schick 2020; Rini 2020; De Ruiter 2021) to a more pragmatic view where humans are expected to adapt to this new situation (e.g., Van Doorn et al. 2021; Harris 2021). It is a reality though, that deepfakes will become more and more a part of the fabric of our daily lives in the (near) future and more academic research towards the impact of this is absolutely required. In this thesis I have researched the impact of deepfakes on morality in the context of extreme beliefs, which led to the following research question: What moral principles are at stake in the use of deepfakes in the context of groups or people holding extreme beliefs in the Benelux in the years 2020, 2021 and ten years in the future?

To answer this question, I have used the reflective equilibrium method to explore various ethical issues on three selected deepfake case studies. The first case study is about an alleged deepfake application of the impersonation of a political person (Navalny) or his representative (Volkov) against a tense geo-political backdrop. The principles that are at stake in this case study is Volkov being wronged as a speaker and the non-consensual use of his persona, the sovereignty of a nation state and the use of satire/ freedom of speech. The outcome of the ethical evaluation for this case study is that there are very few situations where it is acceptable to non-consensually use a person's persona in the context of satire or freedom of speech. In addition to this it is important to evaluate the case study against the geopolitical context. In this case study the Russian comedians claim it was satire, one could argue against the tense geopolitical backdrop that from a Western point of view this would only exacerbate things.

The second case study is about XR Belgium that created a political deepfake in which Belgian prime-minister Sophie Wilmès claimed that the COVID-19 pandemic was the consequence of the looming climate crisis. The ethical issue that I explored in this case study is whether the moral principles governing XR justify the use of deepfakes and who is responsible for the potential harm being done. XR's moral principles in this

case study were principles from climate ethics,[110] combined with the *pro tanto* right to freedom of speech that need to be balanced with the *pro tanto* obligation to do no harm. The direct harm is being done to Wilmès which is governed by the principles of informed consent and right to your own persona. Indirect harm is being done to a part of the Belgian population who may be harmed in their credulity. The result of the reflective equilibrium process revealed that *in itself* it is morally permissible for political actors to use deepfake technology to pursue political goals, but that in an ethical evaluation all *pro tanto* rights and obligations need to be contextually balanced. As a thinker, after all things considered in this case study, I think it is morally impermissible for XR to use a deepfake because of the non-consensual use of Wilmès's persona and there are other viable political means to pursue their goals. Deepfake events like that happened in this case study foster the public debate about the conditions in which the use of deepfakes is morally acceptable. This debate is progressing as the 2021 Mark Rutte-deepfake in the Netherlands showed where some vindicated the use of deepfakes because the climate crisis is such an urgent problem (De Jong 2021; Mommers & Wijnberg 2021).

The third case study is about the use of deepfakes in 2031 and assesses the moral aspects of a deepfake deployed by an organization holding extreme beliefs in a world where fake and reality can only be distinguished by algorithms. The principles at stake here are transparency, epistemic authority, and credulity.[111] On this account it is obvious that it is morally impermissible to use deepfakes, but the reflective equilibrium process revealed interesting observations that are applicable to future cases that deal with deepfakes in the context of extreme beliefs and beyond. The first observation is that deepfakes are not created in a social vacuum and will be weaponized by groups holding extreme views, just like they have weaponized other technologies in the past (e.g., Dan et al. 2021; Schick 2020; Langguth et al. 2021). They will capitalize on the existing sociological trend of diminishing trust in epistemic authorities (e.g., Harambam 2020) and leverage this to create dis- and misinformation. The second observation is that in a society that is inundated with deepfakes people's credulity will be based more and more on technology like deepfake detectors which might lead to algorithmic akrasia.

Applying the reflective equilibrium method on the three case studies has provided insight in the moral properties and governing principles that are at stake when dealing with deepfake technology. The findings of this research will help academic researchers

[110] These principles are ecological justice, sense of urgency for climate change, sustainability and civil disobedience (see §7.4 for more details).

[111] In addition to the principles of informed consent and right to your own persona which already have been covered in the other case studies.

in extreme beliefs and fundamentalism to embed and describe deepfakes in their future research, not as a new and isolated technical category but as a technology wrapped in a broader, social context in our society that will amplify existing sociological trends. Deepfakes can and will be weaponized by groups holding extreme beliefs and should be contextualized as the latest technology manifestation in the creation of disinformation and propaganda. In general mis- and disinformation will lead to an epistemic deterioration of our information environments (De Ridder 2021) and deepfakes will only accelerate and amplify this and the insights from this research will help academics and ethicists take a broader, more nuanced, and contextualized view to assess the moral impact of deepfakes. Another theoretical ramification of this research is using the reflective equilibrium method in applied ethics. In the academic literature a lot is written about the merits and disadvantages of reflective equilibrium as an ethical method, but it was hard to find literature that described in detail *how* the method was applied in concrete cases. For this research I had to develop my own version of the reflective equilibrium method which is a simplified version of the method described by Knight (2017) which can be a suitable method for technology ethics A lesson learned for researchers and ethicists who want to use the reflective equilibrium method is, that you need to adapt the method towards the context of your research.

Lessons learned

Since deepfakes are expected to have a profound impact on our future society, it is essential to have an informed public debate about this topic and the lessons learned, documented in this thesis, can contribute to this. The first lesson learned is that in order to have a fruitful informed discussion about the moral aspects of deepfakes in our society it is important that the population gets educated and informed about what it is, its consequences and how to go about this. This will be an ongoing effort and my recommendation for policy makers is to pay sufficient attention to this phenomenon in the context of media literacy. The second lesson learned is that deepfakes will most likely always involve non-consensual use of someone's persona. My recommendation for policy makers is to increase awareness for this and to improve legislation so that it will be easier to fight this when someone is being non-consensually wronged by deepfakes. The third lesson learned is that the impact of using reality altering technology is similar, regardless whether this is a deepfake, an alleged deepfake or a cheapfake. In future cases where reality is altered by technology, e.g., in the form of satirical impersonation or a deepfake, it important to properly contextualize the situation in terms of geopolitical, cultural, and testimonial background before it is interpreted and morally assessed. My recommendation for policy makers is to look at deepfakes through this lens when creating policies.

Limitations current research and future research avenues

One limitation of the current research is that its design and research question are *not* intended to provide normative answers for deepfake related matters but is intended to provide an overview of the moral principles underpinning the use of deepfakes in the context of extreme beliefs. It would be impossible to come up with any normative statements because deepfakes are a nascent phenomenon and there have not been any documented real-life cases where deepfakes are used by groups holding extreme beliefs. It is expected that deepfakes will become more prevalent in the future and will be weaponized by these groups for their political goals. The insights from this thesis will help inform future research towards the use of technology in the context of extreme beliefs in general. Another limitation is that the technological state of affairs in the future will be assessed through the lens of the state of affairs in 2021. Since ethics is based on an ongoing public debate the insights from this research will be helpful to inform this. There are many interesting avenues for future work that can build on this research, but I will provide four suggestions for this.

1. Deepfakes can be considered as the next category in creating disinformation. Further research could investigate what lessons learned from combatting fake news in the context of groups holding extreme beliefs could be applied to deepfakes, once they start to proliferate.
2. The reflective equilibrium method is a useful method for conducting an ethical assessment in the field of extreme beliefs and technology ethics. However, the method in the academic literature is too abstract to be applied. Further research could be done towards developing an ethical framework that is based on reflective equilibrium that can be applied by ethicists and extreme beliefs researchers when conducting ethical assessments.
3. It is not always necessary to develop new legislation for deepfakes, as it turns out there is room for governing deepfakes in existing laws. Further research should be done to map the current state of legislation and the potential gaps regarding deepfakes in the context of extreme beliefs.
4. The insights of the research in this thesis are focused on countries having a Western worldview. Further research should be done towards the ethics and deepfakes in other cultures or countries having other worldviews.

What lies ahead

The research in this thesis focused on deepfakes, however deepfakes are just one of the technological developments that is capitalizing on the advent of AI in our society. The *Netherlands Scientific Council for Government Policy* (WRR)[112] published a

[112] In Dutch this is *Wetenschappelijke Raad voor Regeringsbeleid* often abbreviated as WRR.

report *Mission AI. The New System Technology* (WRR 2021) in which they argue that AI is a *system technology*, like e.g., the combustion engine and the computer, that is expected to have a big impact on the economy, our society, and our public values. In this report WRR argues, among other advice, that in our society AI needs to be *demystified*, i.e., painting a realistic picture to the Dutch population what AI is, what AI can do and what its limitations are, so a fruitful public debate can take place about the terms and conditions in which our society wants to adopt this technology. Against this backdrop my research in this thesis will be helpful to *demystify* what deepfakes are and will help to push the public debate forward to discuss the moral implication deepfakes will have on our society.

In sum, technology will become more and more important in the way humans perceive the world and interact with the world. It is of utmost importance for researchers in the domain of extreme beliefs, ethics, and fundamentalism to have a good conceptual understanding of technological developments that are shaping our society, not as a technological phenomenon but to see technology as a manifestation and amplifier of underlying sociological trends. I hope this research has helped to demystify the concept of deepfakes and what the moral implications they may invoke. As deepfakes is a relatively new phenomenon but that is expected to have big impact further research in the future is needed to explore the moral and epistemological implications of this.

Bibliography

Aithos. (2019). *ETHICS IN SYNTHETIC MEDIA a guide to building mindful technology* (Boston MA: The AITHOS Coalition). https://www.aithos.technology/aithos-guide.

Ajder, H., Patrini, G., Cavalli, F. & Cullen, L. (2019). *The State of Deepfakes: Landscape, Threats, and Impact* (Amsterdam: Deeptrace Labs). https://sensity.ai/reports/.

Alcántara-Ayala, I., Burton, I., Lavell, A., Mansilla, E., Maskrey, A., Oliver-Smith, A., & Ramírez-Gómez, F. (2021). "Editorial: Root causes and policy dilemmas of the COVID-19 pandemic global disaster." *International Journal of Disaster Risk Reduction*, 52, 101892.

Allen, Anita. (2021). "Privacy and Medicine." *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). https://plato.stanford.edu/entries/privacy-medicine/.

Allcott, Hunt, & Gentzkow, Matthew. 2017. "Social Media and Fake News in The 2016 Election." *Journal of Economic Perspectives* 31 (2), 211–236.

Anderson, Clifford. (2019). "A New Hermeneutics of Suspicion? The Challenge of deepfakes to Theological Epistemology." *Cursor_ Zeitschrift für Explorative Theologie*. https://cursor.pubpub.org/pub/andersondeepfakes/release/8.

Anderson, Clifford. (2021). "Empathy in an Age of Deepfakes." *Cursor_ Zeitschrift für Explorative Theologie*. https://cursor.pubpub.org/pub/anderson-empathy-deepfakes/release/1.

Arras, John. (2009). "The Way We Reason Now: Reflective Equilibrium in Bioethics." In Bonnie Steinbock (ed.), *The Oxford Handbook of Bioethics* (Oxford: Oxford University Press), 46-71.

Ashford, Nicholas. (2021). "Not on Facebook? You're Still Likely Being Fed Misinformation." *New York Times* 29 March 2021. https://www.nytimes.com/2021/03/29/opinion/misinformation-television-radio.html.

Ashton, John. (2021). "COVID-19 and the anti-vaxxers." *Journal of the Royal Society of Medicine*, 114(1), 42–43.

Asscher, L. F. (2002). *Communicatiegrondrechten: een onderzoek naar de constitutionele bescherming van het recht op vrijheid van meningsuiting en het communicatiegeheim in de informatiesamenleving* (Amsterdam: Otto Cramwinckel).

Aupers, Stef, & Harambam, Jaron. (2018). "Rational Enchantments: Conspiracy Theory between Secular Scepticism and Spiritual Salvation." In Asbjørn Dyrendal, David G. Robertson, and Egil Asprem (eds.), *Handbook of Conspiracy Theory and Contemporary Religion* (Leiden: Brill), 48-69.

Balmforth, Tom, & Zverev, Anton. (2021). "Russia hits Navalny with new charge that could add to jail term." *Reuters* 11 August 2021. https://www.reuters.com/world/europe/navalny-faces-new-criminal-charges-over-anti-corruption-foundation-say-russian-2021-08-11/.

Barari, S., Lucas, C., & Munger, K. (2021). "Political Deepfake Videos Misinform the Public, but No More Than Other Fake Media." *OSF Preprints* 13 January 2021.

Baudrillard, Jean. (1994). *Simulacra and Simulation,* translated by Sheila Faria Glaser (Ann Arbor MI: University of Michigan Press).

Baun, C., Kunze, M., Nimis, J., & Tai, S. (2011). *Cloud Computing : Web-Based Dynamic IT Services* (Heidelberg: Springer).

Baurmann, Michael & Cohnitz, Daniel. (2021). "Trust No One? - The (Social) Epistemological Consequences of Belief in Conspiracy Theories." In Sven Bernecker, Amy Floweree and Thomas Grundmann (eds.), *The Epistemology of Fake News* (Oxford UK: Oxford University Press), 334-357.

Beauchamp, Tom. (2016). "The Role of Principles in Practical Ethics." In J. Boyle and L. Sumner (eds.), *Philosophical Perspectives on Bioethics* (Toronto: University of Toronto Press), 79-95.

Belgium. (2016). "Freedom of expression, including freedom of the press." *Kingdom of Belgium Foreign Affairs, Foreign Trade and Development Cooperation*. https://diplomatie.belgium.be/en/policy/policy_areas/human_rights/specific_themes/freedom_expression_including_freedom_press.

Berger, J. M. (2018). *Extremism* (Cambridge MA: The MIT Press).

Blackburn, Simon. (2008). "value." *The Oxford Dictionary of Philosophy*. https://www.oxfordreference.com/view/10.1093/acref/9780199541430.001.0001/acref-9780199541430-e-3225?rskey=gz8O8v&result=3223.

Blaskiewicz, Robert. (2013). “The big Pharma conspiracy theory.” *Medical Writing*, 22(4), 259-261.

Borràs, Susana. (2016). “New Transitions from Human Rights to the Environment to the Rights of Nature.” *Transnational Environmental Law*, 5(1), 113-143.

Bredvold, Louis I. (1940). “A note in defence of satire.” *Elh*, 7(4), 253–264.

Breland, Ali. (2019). “The Bizarre and Terrifying Case of the “Deepfake” Video that Helped Bring an African Nation to the Brink.” *Mother Jones* 15 March 2019. https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

Brewster, Thomas. (2021). “Fraudsters Cloned Company Director’s Voice In $35 Million Bank Heist, Police Find.” *Forbes* 14 October 2021. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=51bf7ef27559.

Bringsjord, Selmer & Govindarajulu, Naveen Sundar. (2018). "Artificial Intelligence." *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). https://plato.stanford.edu/entries/artificial-intelligence/.

Brooks, C. F. (2021). “Popular discourse around deepfakes and the interdisciplinary challenge of fake video distribution.” *Cyberpsychology, Behavior, and Social Networking,* 24(3), 159–163.

Brouwers, Arnout. (2021). “‘Nederlands gesprek met fake-Navalnyteam onderdeel van grote Russische desinformatiecampagne’.” *de Volkskrant* 25 april 2021. https://www.volkskrant.nl/nieuws-achtergrond/nederlands-gesprek-met-fake-navalnyteam-onderdeel-van-grote-russische-desinformatiecampagne~b7b1e3f3/.

Brouwers, Arnout, & Verhagen, Laurens. (2021). “Deepfake? Nee, gewoon een slecht geschminkte grapjurk in een badjas.” *de Volkskrant* 28 May 2021. https://www.volkskrant.nl/nieuws-achtergrond/deepfake-nee-gewoon-een-slecht-geschminkte-grapjurk-in-een-badjas~b2866b50/.

Brown, James Robert, & Fehige, Yiftach. (2019). "Thought Experiments." *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2019/entries/thought-experiment/.

Brun, Georg. (2017). "Thought Experiments in Ethics." In Michael T. Stuart , Yiftach Fehige and James Robert Brown (eds.), *The Routledge Companion to Thought Experiments* (Abingdon UK: Routledge), 195-210.

Bortolotti, Lisa, & Ichino, Anna. (2020). "Conspiracy theories may seem irrational – but they fulfill a basic human need." *The Conversation* 9 December 2020. https://theconversation.com/conspiracy-theories-may-seem-irrational-but-they-fulfill-a-basic-human-need-151324.

Cassam, Quassim. (2019a). *Conspiracy Theories* [Kindle version] (Cambridge UK: Polity Press).

Cassam, Quassim. (2019b). *Vices of the mind : from the intellectual to the political* (Oxford UK: Oxford University Press).

Cassam, Quassim. (2021). *Extremism : A Philosophical Analysis* (Abingdon UK: Routledge).

Cath, Yuri. (2016). "Reflective Equilibrium." In Herman Cappelen, Tamar Szabó Gendler & John Hawthorne (eds.), *The Oxford Handbook of Philosophical Methodology* (Oxford: Oxford University Press), 213-230.

Cheng, Sarah, & Norcross, Harriet. (2021) "Internet Censorship in the Time of a Global Pandemic: A Proposal for Revisions to Section 230 of the Communications Decency Act," *Brigham Young University Prelaw Review*, Vol. 35 , Article 11.

Chesney, Bobby, & Citron, Danielle. (2019). "Deep fakes: A looming challenge for privacy, democracy, and national security." *California Law Review*, 107, 1753-1820.

Chignell, Andrew. (2018). "The Ethics of Belief." *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). https://plato.stanford.edu/entries/ethics-belief/.

Christman, John. (2020). "Autonomy in Moral and Political Philosophy." *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). https://plato.stanford.edu/entries/autonomy-moral/.

Coeckelbergh, Mark. (2009). "Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents." *AI & SOCIETY*, 24(2), 181-189.

CognitionX. (2021). “Cyber & Defence: Will 2030 be real? A discussion on Deepfakes.” *YouTube* 15 June 2021. https://youtu.be/c_f6Xuqtlus.

Cragg, Wesley. (2000). “Human rights and business ethics: fashioning a new social contract.” *Journal of Business Ethics*, 27(1-2), 205–214.

Culkin, John. (1967). “A schoolman’s guide to Marshall McLuhan.” *Saturday Review* 18 March 1967, 51-53, 70-72.

Dale, R. (2021). “GPT-3: What’s it good for?” *Natural Language Engineering*, 27(1), 113-118.

Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021). “Visual Mis- and Disinformation, Social Media, and Democracy.” *Journalism & Mass Communication Quarterly* 98(3), 641–664.

Dancy, Jonathan. (2021).”The Role of Imaginary Cases in Ethics.” In *Practical Thought: Essays on Reason, Intuition, and Action* (Oxford: Oxford University Press), 62-74.

Daniels, Norman. (1996). *Justice and justification: Reflective equilibrium in theory and practice* (New York: Cambridge University Press).

Daniels, Norman. (2016). "Reflective Equilibrium." *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). https://plato.stanford.edu/archives/sum2020/entries/reflective-equilibrium/.

De Graaf, Beatrice. (2021). *Radicale verlossing : wat terroristen geloven* [Kindle version] (Amsterdam: Prometheus).

De Graaf, Beatrice, & Van den Bos, Kees. (2021). “Religious Radicalization: Social Appraisals and Finding Radical Redemption in Extreme Beliefs.” *Current Opinion in Psychology* 40, 56–60.

De Jong, Addy. (2021). “Nepfilmpje Rutte sterk staaltje communicatie.” *Reformatorisch Dagblad* 30 October 2021. https://www.rd.nl/artikel/948784-nepfilmpje-rutte-sterk-staaltje-communicatie.

De Jonge, Manon. (2021). *From Mont-Sainte Victoire to Mont-S'AI'nte Victoire: Cézanne through the lens of Generative Adversarial Networks*, Bachelor thesis Arts & Culture Studies Radboud University. https://theses.ubn.ru.nl/handle/123456789/11265.

De Maagt, Sem. (2017). “Reflective equilibrium and moral objectivity.” *Inquiry*, 60 (5), 443-465.

De Ridder, Jeroen. (2021). “What's so bad about misinformation?” *Inquiry*, 1-23.

De Ruiter, Adrienne. (2021). “The Distinct Wrong of Deepfakes.” *Philosophy & Technology*, 1-22.

De Standaard. (2020). “Klimaatactivisten manipuleren speech Wilmès.” *De Standaard* 14 April 2020. https://www.standaard.be/cnt/dmf20200414_04922023.

Dentith, Matthew. (2014). *The Philosophy of Conspiracy Theories* (Basingstoke UK: Palgrave Macmillan).

Dever, Josh. (2016). “What is Philosophical Methodology?” In Herman Cappelen, Tamar Szabó Gendler & John Hawthorne (eds.), *The Oxford Handbook of Philosophical Methodology* (Oxford: Oxford University Press), 3-24.

Diakopoulos, Nicholas, & Johnson, Deborah. (2020). “Anticipating and addressing the ethical implications of deepfakes in the context of elections.” *New Media & Society*, 23(7), 2072–2098.

Diep Nep. (2021). “This is not Morgan Freeman - A Deepfake Singularity.” *YouTube* 7 July 2021. https://youtu.be/oxXpB9pSETo.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). “Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?” *The International Journal of Press/Politics*, 26(1), 69–91.

Dodd, Vikram, & Grierson, Jamie. (2020). “Terrorism police list Extinction Rebellion as extremist ideology.” *The Guardian* 10 January 2020. https://www.theguardian.com/uk-news/2020/jan/10/xr-extinction-rebellion-listed-extremist-ideology-police-prevent-scheme-guidance.

Doorn, Neelke. (2010). “Applying Rawlsian Approaches to Resolve Ethical Issues: Inventory and Setting of a Research Agenda.” *Journal of Business Ethics* 91, no. 1, 127–143.

Doorn, Neelke. (2012). “Exploring Responsibility Rationales in Research and Development (R&D).” *Science, Technology, & Human Values* 37, no. 3, 180–209.

Doorn, N., & Taebi, B. (2018). “Rawls's Wide Reflective Equilibrium As a Method for Engaged Interdisciplinary Collaboration: Potentials and Limitations for the Context of Technological Risks.” *Science, Technology & Human Values* 43, no. 3, 487–517.

Dorobanțu, Marius. (2020). *Theological Anthropology and the Possibility of Human-Level Artificial Intelligence: Rethinking Human Distinctiveness and the Imago Dei,* doctoral dissertation (Strassbourg: University of Strassbourg).

Drążkiewicz Grodzicka, Elżbieta, & Harambam, Jaron. (2021). “What should academics do about conspiracy theories? Moving beyond debunking to better deal with conspiratorial movements, misinformation and post-truth.” *Journal for Cultural Research*, 25(1), 1-11.

Dutilh Novaes, Catarina, & De Ridder, Jeroen. (2021). “Is Fake News Old News?” In Sven Bernecker, Amy Floweree and Thomas Grundmann (eds.), *The Epistemology of Fake News* (Oxford UK: Oxford University Press), 156-179.

EU FRA. (2021). “EU Charter of Fundamental Rights - Freedoms.” *EU Agency for Fundamental Rights* accessed 28 September 2021. https://fra.europa.eu/en/eu-charter/title/title-ii-freedoms.

Extinction Rebellion. (2020a). “#TELLTHETRUTHBELGIUM The Truth About COVID-19 And The Ecological Crisis - A Speech For Sophie Wilmès.” https://www.extinctionrebellion.be/en/tell-the-truth.

Extinction Rebellion. (2020b). “The Prime Minister's Speech by Our Rebels.” https://www.extinctionrebellion.be/en/tell-the-truth/the-prime-ministers-speech-by-the-rebels.

Extinction Rebellion. (2021a). “We demand.” Accessed 26 October 2021. https://www.extinctionrebellion.be/en/.

Extinction Rebellion (2021b). “What is Climate and Ecological Justice?” Accessed 29 October 2021. https://extinctionrebellion.uk/the-truth/about-us/what-is-climate-and-ecological-justice/.

Extreme Beliefs. (2021). “Extreme beliefs, the epistemology and ethics of fundamentalism.” *Extreme Beliefs*. https://extremebeliefs.com/.

Fallis, Don. (2020). “The epistemic threat of deepfakes.” *Philosophy & Technology*, 1-21.

FBI. (2021). “Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations.” *FBI Private Industry Notification* 210310-001 10 March 2021. https://www.ic3.gov/Media/News/2021/210310-2.pdf.

Felix Meritis. (2020). “Future Affairs N°1 | De Infocalyps (met o.a. Sebastiaan van der Lans en Catarina Dutilh Novaes).” *YouTube* 18 December 2020. https://youtu.be/xWrrGnwq1gE.

Fieser, James. (2021). “Ethics.” *Internet Encyclopedia of Philosophy*. Accessed 3 September 2021. https://iep.utm.edu/ethics/.

Fletcher, John. (2018). “Deepfakes, artificial intelligence, and some kind of dystopia: the new faces of online post-fact performance.” *Theatre Journal*, 70(4), 455–471.

Floridi, Luciano. (2013). “Distributed morality in an information society.” *Science and Engineering Ethics*, 19(3), 727-743.

Floridi, Luciano. (2018). “Artificial intelligence, deepfakes and a future of ectypes.” *Philosophy & Technology*, 31(3), 317-321.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing* (New York: Oxford University Press).

Galindo, Gabriela. (2020). “XR Belgium posts deepfake of Belgian premier linking COVID-19 with climate crisis.” *The Brussels Times* 14 April 2020. https://www.brusselstimes.com/news/belgium-all-news/politics/106320/xr-belgium-posts-deepfake-of-belgian-premier-linking-COVID-19-with-climate-crisis/.

Gardiner, Stephen. (2010). “A Perfect Moral Storm: Climate Change, Intergenerational Ethics, and the Problem of Corruption.” In Stephen M. Gardiner, Simon Caney, Dale Jamieson, & Henry Shue (eds.), *Climate Ethics : Essential Readings* (New York: Oxford University Press), 87-98.

Climate Ethics (p. iii). Oxford University Press. Kindle Edition.

Climate Ethics (p. 87). Oxford University Press. Kindle Edition.

Gellers, Joshua., (2020). *Rights for Robots: Artificial Intelligence, Animal and Environmental Law* (London UK: Routledge).

Giansiracusa, Noah. (2021). *How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More* (Acton MA: Appress).

Goodfellow, I., Pouget-Abadi, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). "Generative Adversarial Nets." *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)* 2: 2672–2680.

Gordon, John-Stewart, & Nyholm, Sven. (2021). "Ethics of Artificial Intelligence." *Internet Encyclopedia of Philosophy*. Accessed 5 October 2021. https://iep.utm.edu/ethic-ai/.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). "Algorithmic content moderation: Technical and political challenges in the automation of platform governance." *Big Data & Society*.

Greengard, Samuel. (2020). "Will deepfakes do deep damage?" *Communications of the ACM* 63, no. 1, 17-19.

Gregory, Sam. (2021). "Liar's dividend alert." *Twitter* 1 October 2021. https://twitter.com/SamGregory/status/1443979471158226945.

Griffin, James. (1993). "How We Do Ethics Now." *Royal Institute of Philosophy Supplement* 35, 159–177.

Grundmann, Reiner. (2017). "The Problem of Expertise in Knowledge Societies." *Minerva* 55, 25–48.

Hancock, Jeffrey, & Bailenson, Jeremy. (2021). "The social impact of deepfakes." *Cyberpsychology, Behavior, and Social Networking*, 24(3), 149–152.

Hao, Karen. (2021). "How Facebook got addicted to spreading misinformation." *MIT Technology Review* 11 March 2021. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

Harambam, Jaron, & Aupers, Stef. (2019). "From the unbelievable to the undeniable: Epistemological pluralism, or how conspiracy theorists legitimate their extraordinary truth claims." *European Journal of Cultural Studies*, 24(4), 990–1008.

Harambam, Jaron. (2020). *Contemporary Conspiracy Culture: Truth and Knowledge in an Era of Epistemic Instability* (Abingdon UK: Routledge).

Harambam, Jaron. (2021). "Against modernist illusions: why we need more democratic and constructivist alternatives to debunking conspiracy theories." *Journal for Cultural Research*, 25(1), 104-122.

Harris, Keith Raymond. (2021). "Video on demand: what deepfakes do and how they harm." *Synthese*.

Hauser, Larry. (2021). "Artificial Intelligence." *Internet Encyclopedia of Philosophy*. Accessed 5 October 2021. https://iep.utm.edu/art-inte/.

Hetmański, Marek. (2020). "Expertise and Expert Knowledge in Social and Procedural Entanglement." *Eidos: A Journal for Philosophy of Culture* 4, no. 2, 6-22.

Hilary, I. O., & Dumebi, O.-O. (2021). "Social Media as a Tool for Misinformation and Disinformation Management." *Linguistics and Culture Review*, 5(S1), 496-505.

Holubowicz, Gerald. (2020). "Extinction Rebellion s'empare des deepfakes en Belgique." *Mediapart* 18 April 2020. https://blogs.mediapart.fr/geraldholubowicz/blog/150420/extinction-rebellion-s-empare-des-deepfakes-en-belgique.

Ichino, Anna, & Räikkä, Juha. (2020). "Non-doxastic conspiracy theories." *Argumenta*.

Jasper, David. (2004). *A Short Introduction to Hermeneutics* (Louisville KY: Westminster John Knox).

Johnson, Deborah, & Diakopoulos, Nicholas. (2021). "What to do about deepfakes." *Communications of the ACM*, 64(3), 33-35.

Jones, Jonathan. (2018). "A portrait created by AI just sold for $432,000. But is it really art?" *the Guardian* 26 October 2021. https://www.theguardian.com/artanddesign/shortcuts/2018/oct/26/call-that-art-can-a-computer-be-a-painter.

Jutzi, C. A., Willardt, R., Schmid, P. C., & Jonas, E. (2020). "Between Conspiracy Beliefs, Ingroup Bias, and System Justification: How People Use Defense Strategies to Cope With the Threat of COVID-19." *Frontiers in Psychology*, 11 (2538).

Kahneman, Daniel. (2011). *Thinking, Fast and Slow* (London UK: Penguin Books).

Katsafanas, Paul. (2020). “Group Fanaticism and Narratives of Ressentiment.” In Michael Staudigl, Hans Bernard Schmid, Ruth Tietjen & Leo Townsend (eds.), *Confronting Fanaticism* (Abingdon UK: Routledge), forthcoming.

Kelly, Thomas, & McGrath, Sarah. (2010). “Is Reflective Equilibrium Enough?” *Philosophical Perspectives* 24, no. 1, 325–359.

Kerner, Catherine, & Risse, Mathias. (2021). “Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds.” *Moral Philosophy and Politics,* 8(1), 81-108.

Kessler, F., & Schäfer, M.T. (2018). “Trust in techno-images: Early media collections as precursors of big data.” *TMG Journal for Media History*, *21*(2).

Knight, C. (2017). “Reflective Equilibrium.” In A. Blau (ed.), *Methods in Analytical Political Theory* (Cambridge UK: Cambridge University Press), 46–64.

Kwok, Andrei, & Koh, Sharon. (2021). “Deepfake: a social construction of technology perspective.” *Current Issues in Tourism* 24:13, 1798-1802.

Langa, Jack. (2021). “Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes.” *Boston University Law Review*, 101(2), 761-802.

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P., & Schroeder, D.T. (2021). “Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes.” *Frontiers in Communication*, 6 (26).

Lanham, Michael. (2021). “GANs, GANs, and More GANs.” In: *Generating a New Reality* (Berkeley CA: Apress), 105-134.

Lomsadze, Giorgi. (2021). “Georgia’s big little election.” *Eurasianet* 1 October 2021. https://eurasianet.org/georgias-big-little-election.

Mahan, Logan. (2021). “A New Deepfake Technology Allows You to Make People in Photographs Say Whatever You Want.” *InsideHook* 4 October 2021. https://www.insidehook.com/daily_brief/internet/deepfake-technology-say-anything.

Malaria Must Die. (2019). “David Beckham speaks nine languages to launch Malaria Must Die Voice Petition.” *YouTube* 9 April 2019. https://youtu.be/QiiSAvKJIHo.

Maras, Marie-Helen, & Alexandrou, Alex. (2019). “Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos.” *The International Journal of Evidence & Proof*, 23(3), 255–262.

McCauley, Clark & Moskalenko, Sophia. (2017). “Understanding Political Radicalization: The Two-Pyramids Model.” *American Psychologist* 72(3), 205-216.

McGuffie, Kris, & Newhouse, Alex. (2020). “The radicalization risks of GPT-3 and advanced neural language models.” *arXiv preprint* 2009.06807. https://arxiv.org/abs/2009.06807.

McKinlay, Roger. (2016). “Technology: Use or lose our navigation skills.” *Nature* 531, 573–575.

Meckel, Miriam, & Steinacker, Léa. (2021). “Hybrid Reality: The Rise of Deepfakes and Diverging Truths.” *Morals & Machines*, 1(1), 10-21.

Molinari, Giulia. (2021). *Is civil society activism necessary to defeat climate change? A reading through the lens of the movement Extinction Rebellion*. Bachelor Thesis Luiss University Rome. http://tesi.luiss.it/30230/.

Mommers, Jelmer, & Wijnberg, Rob. (2021). “Verantwoording: Hoe onze deepfake klimaattoespraak van Mark Rutte tot stand kwam (en waar die op is gebaseerd).” *De Correspondent* 28 October 2021. https://decorrespondent.nl/12846/verantwoording-hoe-onze-deepfake-klimaattoespraak-van-mark-rutte-tot-stand-kwam-en-waar-die-op-is-gebaseerd/1415744814-4527c8cb.

Moreland, J. P., & Craig, W. L. (2017). *Philosophical foundations for a Christian worldview* (Downers Grove IL: InterVarsity Press).

Moriarty, Jeffrey. (2021). “Business Ethics.” *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.), https://plato.stanford.edu/entries/ethics-business/.

Morris, James. (2021). “Simulacra in the Age of Social Media: Baudrillard as the Prophet of Fake News.” *Journal of Communication Inquiry,* 45(4), 319–336.

Moscow Times, The. (2021). “‘Deepfake’ Navalny Aide Targets European Lawmakers.” *The Moscow Times* 23 April 2021. https://www.themoscowtimes.com/2021/04/23/deepfake-navalny-aide-targets-european-lawmakers-a73717.

Moyaert, Marianne. (2014). *In Response to the Religious Other : Ricoeur and the Fragility of Interreligious Encounters* (Lanham ML: Lexington Books).

Moyaert, Marianne. (2019). "Interreligious Hermeneutics, Prejudice, and the Problem of Testimonial Injustice." *Religious Education*, 114(5), 609-623.

Napolitano, Giulia. (2021). "Conspiracy Theories and Evidential Self-Insulation." In Sven Bernecker, Amy Floweree and Thomas Grundmann (eds.), *The Epistemology of Fake News* (Oxford UK: Oxford University Press), 82-105.

NCTV. (2021). "Definities gebruikt in het Dreigingsbeeld Terrorisme Nederland." *Nationaal Coördinator Terrorismebestrijding en Veiligheid*, accessed 15 October 2021. https://www.nctv.nl/onderwerpen/dtn/definities-gebruikt-in-het-dtn.

Nguyen, C.T. (2020). "Echo Chambers and Epistemic Bubbles." *Episteme* 17, no. 2, 141–161.

Nissenbaum, Helen. (2011). "A contextual approach to privacy online." *Daedalus*, 140(4), 32–48.

NOS Nieuws. (2021a). "Stafleider Navalny na nepgesprek Kamer: 'Kremlin gebruikt Zoom als wapen'." *NOS Nieuws* 25 April 2021. https://nos.nl/artikel/2378207-stafleider-navalny-na-nepgesprek-kamer-kremlin-gebruikt-zoom-als-wapen.

NOS Nieuws. (2021b). "Kamervoorzitter: gesprek met nep-Volkov had voorkomen kunnen worden." *NOS Nieuws* 28 May 2021. https://nos.nl/artikel/2382643-kamervoorzitter-gesprek-met-nep-volkov-had-voorkomen-kunnen-worden.

NOS Nieuws. (2021c). "Tientallen doden in Duitsland door watersnood, ook in België doden." *NOS Nieuws* 15 July 2021. https://nos.nl/collectie/13869/artikel/2389386-tientallen-doden-in-duitsland-door-watersnood-ook-in-belgie-doden.

Nozick, Robert. (1997). "The Characteristic Features of Extremism." In *Socratic Puzzles* (Cambridge MA: Harvard University Press), 296-299.

Obvious. (2018). "Obvious, explained." *Medium* 14 February 2018. https://medium.com/@hello.obvious/aithe-rise-of-a-new-art-movement-f6efe0a51f2e.

Paris, Britt, & Donovan, Joan. (2019). "Deepfakes and cheap fakes." *Data & Society*. https://datasociety.net/library/deepfakes-and-cheap-fakes/.

Peels, Rik. (2011). “Tracing culpable ignorance.” *Logos & Episteme*, 2(4), 575-582.

Peels, Rik. (2020). *Defining ‘Fundamentalism’*, Manuscript submitted for publication.

Peels, Rik. (2021). “Responsibility for fundamentalist belief.” In Kevin McCain & Scott Stapleford (eds.), *Epistemic Duties: New Arguments, New Angles* (New York: Routledge), 221-238.

Philpott, Daniel. (2020). "Sovereignty." *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2020/entries/sovereignty.

Pierre, Joseph. (2020). “Mistrust and misinformation: a two-component, socio-epistemic model of belief in conspiracy theories.” *Journal of Social and Political Psychology*, 8(2), 617–641.

Rawls, John. (1971). *A Theory of Justice* (Cambridge MA: Belknap Press of Harvard University Press).

Rawls, John. (1974). “The independence of moral theory.” *Proceedings and Addresses of the American Philosophical Association*, 48, 5–22.

Ray, Andrew. (2021). ”Disinformation, deepfakes and democracies: The need for legislative reform.” *The University of New South Wales Law Journ*al, 44(3), 983–1013.

Reisner, Andrew. (2013). “Prima Facie and Pro Tanto Oughts.” In Hugh Lafollette (ed.), *International Encyclopedia of Ethics* (Hoboken NJ: Wiley). https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444367072.wbiee406.

Rini, Regina. (2020). "Deepfakes and the Epistemic Backstop." *Philosophers' Imprint* 20, no.24, 1–16.

Rini, Regina, & Cohen, Leah. (2021). “Deepfakes, Deep Harms.” *Journal of Ethics and Social Philosophy*, forthcoming.

Rivera-Ferre, M. G., López-i-Gelats Feliu, Ravera, F., Oteros-Rozas, E., di Masso, M., Binimelis, R., & El Bilali, H. (2021). “The two-way relationship between food systems and the COVID19 pandemic: causes and consequences.” *Agricultural Systems*, 191.

Roth, Andrew. (2021). “European MPs targeted by deepfake video calls imitating Russian opposition.” *the Guardian* 22 April 2021. https://www.theguardian.com/world/2021/apr/22/european-mps-targeted-by-deepfake-video-calls-imitating-russian-opposition.

Ross, W. D., & Stratton-Lake, P. (2002). *The right and the good* (Oxford: Oxfurf University Press).

Rotter, Julian. (1980). “Interpersonal trust, trustworthiness, and gullibility.” *American psychologist*, 35(1), 1.

Rottman, J., Crimston, C. R., & Syropoulos, S. (2021). “Tree-Huggers Versus Human-Lovers: Anthropomorphism and Dehumanization Predict Valuing Nature Over Outgroups.” *Cognitive Science*, 45(4), e12967.

Rubiano A., Maria Paula. (2021). “Navigating Qualitative Research.” *The Open Notebook 7* September 2021. https://www.theopennotebook.com/2021/09/07/navigating-qualitative-research/.

Russell, Stuart, & Norvig, Peter. (2021). *Artificial Intelligence: A Modern Approach* 4th edition (London: Pearson Education).

Savin, Andrej. (2021). “The EU Digital Services Act: Towards a More Responsible Internet.” *Copenhagen Business School, CBS LAW Research Paper*, (21-04). *Journal of Internet Law* forthcoming.

Schaake, Marietje. (2021). “Deepfake ondermijnt liberale democratie.” *NRC* 30 April 2021. https://www.nrc.nl/nieuws/2021/04/30/deepfake-ondermijnt-liberale-democratie-a4041871.

Schick, Nina. (2020). *Deep Fakes and the Infocalypse: What You Urgently Need To Know* [Kindle version] (London UK: Octopus Publishing Group).

Schroten, Egbert. (1998). ”The ‘Herman Case’: The Usefulness of the Wide Reflective Equilibrium model for Ethics Committees.” In W. van der Burg and T. van Willigenburg (eds.), *Reflective equilibrium: Essays in Honour of Robert Heeger* (Dordrecht: Kluwer Academic Publishers), 219–229.

Shah, Deepa. (2019). “Extinction Rebellion: radical or rational?” *British Journal of General Practice*, 69(684), 345-345.

Sheridan, Heather. (2007). “Evaluating technical and technological innovations in sport: Why fair play isn't enough.” *Journal of Sport and Social Issues*, 31 (2), 179-194.

Shevchenko, Vitaly. (2018). “Analysis: Telephone pranksters as a Kremlin media tool.” *BBC Monitoring* 2 January 2018. https://monitoring.bbc.co.uk/product/c1dofhaw.

Singer, Peter. (1981). *The Expanding Circle* (Oxford: Clarendon Press).

Spannring, Reingard. (2021). “Youth in the Anthropocene. Questions of Intergenerational Justice and Learning in a More-Than-Human World.” In Gerald Knapp & Hannes Krall (eds.), Youth *Cultures in a Globalized World* (Cham CH: Springer), 113-133.

Stake, Robert. (2009). “The case study method in social inquiry.” In R. Gomm, M. Hammersley & P. Foster (eds.), *Case study method* (London UK: SAGE Publications), 18-26.

Stephensen, Jan L. (2019). “Towards a Philosophy of Post-creative Practices? – Reading Obvious ‘Portrait of Edmond de Belamy’.” *Politics of the Machine Beirut* 2019 2, 21-30.

Stone, Christopher. (1972). “Should Trees Have Standing--Toward Legal Rights for Natural Objects.” *Southern California Law Review*, 45(2), 450-501.

Stueber, Karsten. (2019). "Empathy", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), https://plato.stanford.edu/entries/empathy/.

Stupp, Catherine. (2019). “Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case.” *The Wall Street Journal* 30 August 2019. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

Swinburne, Richard. (2004*). The Existence of God* (Oxford UK: Oxford University Press).

Synthesia. (2021). “Responsible use of Synthetic Media.” *Ethics*. https://www.synthesia.io/ethics.

Szanto, Thomas. (2020). “Sacralizing Hostility: Fanaticism as a Group-Based Affective Mechanism.” In Michael Staudigl, Hans Bernard Schmid, Ruth Tietjen & Leo

Townsend (eds.), *Confronting Fanaticism* (Abingdon UK: Routledge), forthcoming.

Thalen, Mikael. (2021). “Justin Bieber got duped into picking a fight with a Tom Cruise deepfake.” *daily dot* 8 October 2021. https://www.dailydot.com/debug/justin-bieber-tom-cruise-deepfake/.

Thomas, G., & Myers, K. (2015). *The anatomy of the case study* (London UK: SAGE Publications).

Toews, Rob. (2020). “Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared.” *Forbes* 25 May 2020. https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/?sh=54c134f67494.

Turkle, Sherry. (2021). *The Empathy Diaries : a Memoir* (New York: Penguin Press).

Ume, Chris. (2021). “@deeptomcruise – Making real music again!” *TikTok* 30 August 2021. https://www.tiktok.com/@deeptomcruise/video/7001976758710848774.

Vaccari, Cristian, & Andrew Chadwick. (2020). "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Social Media + Society*.

Van Assen, Mark. (2021). “Russen hielden Tweede Kamer voor de gek met schmink: 'Dit is niet grappig'.” *AD* 4 May 2021. https://www.ad.nl/tech/russen-hielden-tweede-kamer-voor-de-gek-met-schmink-dit-is-niet-grappig~af5991d9/.

Van den Beld, Ton. (1998). ”Background Theories and Religious Beliefs: Their Role and Relation in Reflective Equilibrium” In W. van der Burg and T. van Willigenburg (eds.), *Reflective equilibrium: Essays in Honour of Robert Heeger* (Dordrecht: Kluwer Academic Publishers), 73–88.

Van den Hoven, Jeroen. (1997). “Computer Ethics and Moral Methodology.” *Metaphilosophy* 28, no. 3, 234–248.

Van der Burg, Wibren. (1998). ”Ideals and Ideal Theory: The Problem of Methodological Conservatism.” In W. van der Burg and T. van Willigenburg (eds.), *Reflective equilibrium: Essays in Honour of Robert Heeger* (Dordrecht: Kluwer Academic Publishers), 89–99.

Van der Burg, Wibren, & Van Willigenburg, Theo. (eds.) (1998). *Reflective equilibrium: Essays in Honour of Robert Heeger* (Dordrecht: Kluwer Academic Publishers).

Van der Linden, Sander, & Roozenbeek, Jan. (2020). "Psychological inoculation against fake news." *The psychology of fake news: Accepting, sharing, and correcting misinformation*, 147-169.

Van der Togt, Tony. (2020). "In search of a European Russia strategy." *Atlantisch Perspectief*, 44(1), 36–41.

Van Doorn, M., Duivestein, S. & Pepping, T. (2021). *Real Fake - Playing with Reality in the Age of AI, Deepfakes and the Metaverse* [Kindle Edition] (Voorschoten: Ludibrium [Bot]).[113]

Van Thiel, Ghislaine, & Van Delden, Johannes. (2010). "Reflective Equilibrium As a Normative Empirical Model." *Ethical Perspectives* 17, no. 2, 183–202.

Verbeek, Peter-Paul. (2006). "Persuasive Technology and Moral Responsibility Toward an ethical framework for persuasive technologies." *Persuasive*, 6, 1-15.

Verhagen, Lourens. (2021). "Kamerleden spraken met een nepversie van stafchef van Navalny: deepfake of dubbelganger?" *de Volkskrant* 24 April 2021. https://www.volkskrant.nl/nieuws-achtergrond/kamerleden-spraken-met-een-nepversie-van-stafchef-van-navalny-deepfake-of-dubbelganger~b57ab1ab/

Verweij, Marcel. (1998). "Moral Principles: Authoritative Norms or Flexible Guidelines?" In W. van der Burg and T. van Willigenburg (eds.), *Reflective equilibrium: Essays in Honour of Robert Heeger* (Dordrecht: Kluwer Academic Publishers), 89–99.

Vincent, James. (2021). "'Deepfake' that supposedly fooled European politicians was just a look-alike, say pranksters." *The Verge* 30 April 2021. https://www.theverge.com/2021/4/30/22407264/deepfake-european-polticians-leonid-volkov-vovan-lexus.

Von der Burchard, Hans. (2018). "Belgian Socialist Party Circulates 'Deep Fake' Donald Trump Video." *Politico* 21 May 2018. https://

[113] Ludibrium is not the real name of the publisher; it is a playful name that is based on playing with reality which is one of the key messages of the book. The official name of the publisher is Bot (which does appear on the cover of the Dutch version of the book).

www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/

Vooruit. (2018). “Trump heeft een boodschap voor alle Belgen… #klimaatpetitie.” *Twitter* 20 May 2018. https://twitter.com/vooruit_nu/status/998089909369016325.

Vosoughi, S., Roy, D., & Aral, S. (2018). “The Spread of True and False News Online.” *Science* 359 (6380), 1146−1151.

Vovan222prank. (2021). “Video-prank with the Parliament of the Netherlands (Eng).” *YouTube* 27 May 2021. https://www.youtube.com/watch?v=rmeQkGNO2Zs.

Walker, Shaun. (2016). “Kremlin calling? Meet the Russian pranksters who say 'Elton owes us'.” *The Guardian* 13 March 2016. https://www.theguardian.com/world/2016/mar/13/kremlin-calling-russian-pranksters-elton-john-owes-us.

Wall, Tom. (2021). “Wrong to label Extinction Rebellion as extremists, says Home Office adviser.” *The Guardian* 21 August 2021. https://www.theguardian.com/environment/2021/aug/21/wrong-to-label-extinction-rebellion-as-extremists-says-home-office-adviser.

Walsh, A. (2011). “A Moderate Defence of the Use of Thought Experiments in Applied Ethics.” *Ethical Theory and Moral Practice*, 14(4), 467-481.

Walter, D., Ophir, Y., & Jamieson, K. H. (2020). “Russian Twitter Accounts and the Partisan Polarization of Vaccine Discourse, 2015–2017.” *American Journal of Public Health*, 110(5), 718-724.

Wardle, Claire, & Derakhshan, Hossein. (2017). “Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making – A publication of the Council of Europe.” *Resource Center on Media Freedom in Europe*. https://www.rcmediafreedom.eu/Publications/Reports/Information-disorder-Toward-an-interdisciplinary-framework-for-research-and-policy-making.

Weisberg, Michael. (2016). “Modeling.” In Herman Cappelen, Tamar Szabó Gendler & John Hawthorne (eds.), *The Oxford Handbook of Philosophical Methodology* (Oxford: Oxford University Press), 262-286.

Westerlund, Mika. (2019). “The Emergence of Deepfake Technology: A Review.” *Technology Innovation Management Review* 9, no. 11, 40–53.

Williams, Mary Elizabeth. (2014). “Satire shouldn't come with a warning label.” *Salon* 18 August 2014. https://www.salon.com/2014/08/18/satire_shouldnt_come_with_a_warning_label/.

WITNESS. (2021). “Prepare, Don’t Panic: Synthetic Media and Deepfakes” *WITNESS Media Lab*. https://lab.witness.org/projects/synthetic-media-and-deep-fakes/.

WRR. (2021). “Summary Mission AI. The New System Technology.”[114] The *Netherlands Scientific Council for Government Policy*. https://english.wrr.nl/publications/reports/2021/11/11/summary-mission-ai.

Yadlin-Segal, Aya, & Oppenheim, Yael. (2021). “Whose dystopia is it anyway? Deepfakes and social media regulation.” *Convergence: The International Journal of Research into New Media Technologies*, 27(1), 36–51.

Zuboff, Shoshana. (2019). *The Age of Surveillance Capitalism: the* Fight *for the Future at the New Frontier of Power* (New York: PublicAffairs).

[114] This is the English summary of the full report. Full report is written in Dutch and can be downloaded from the WRR’s website: https://www.wrr.nl/adviesprojecten/artificiele-intelligentie/documenten/rapporten/2021/11/11/opgave-ai-de-nieuwe-systeemtechnologie.

Appendices

In the table below an overview of the appendices used in this thesis can be found.

Appendix	Title
A	Example of animated photo created by *DeepNostalgia*.
B	Picture of havoc wreaked by rainfall in Germany (July 2021).
C	*Migrant Mother*, picture of Dorothea Lange (1936).

Table 11 List of appendices.

Appendix A. Animated photo created by *DeepNostalgia*.

Figure 12 Example of animated photo by DeepNostalgia. The original picture is at the bottom left and a screenshot of the animation is the larger picture top-right (source: https://www.myheritage.nl/deep-nostalgia - screenshot taken on 22 October 2021).

Appendix B. Picture of havoc wreaked by rainfall in Germany (July 2021).

Figure 13 Houses being destroyed in Schuld (Germany) because of floods (source: NOS Nieuws 2021c).

Appendix C. *Migrant Mother*, picture by Dorothea Lange (1936).

Figure 14 'Migrant Mother,' Dorothea Lange's iconic photograph, taken in 1936 during the Great Depression (source: De Lange 2021: 14).