Imagining Deceptive Deepfakes

An ethnographic exploration of fake videos

Tormod Dag Fikse

Master's thesis

ESST – Society, Science and Technology in Europe

University of Oslo

29.10.2018

Imagining Deceptive Deepfakes

An ethnographic exploration of fake videos

2018

Imagining Deceptive Deepfakes: *An ethnographic exploration of fake videos*

Tormod Dag Fikse

http://www.duo.uio.no/

Print: Grafisk Senter

Abstract

In the last year, high quality face-swapped videos have started circulating on the Internet, showing faked situations of politicians doing odd speeches and celebrities performing in pornography. The videos and the surrounding phenomenon have been dubbed *Deepfakes*, alluding to the fact that they are fakes produced by deep learning—a form of machine learning. Commentators have warned that misuse of face-swapped videos in the style of *fake news* could have serious consequences. The technology is part of an idea of a world in which false videos are wreaking havoc. This is one of the complex ways in which the Deepfakes technology is associated with socio-technical practices. This thesis sheds light on these associations through an ethnography of the Internet and autoethnographical accounts of engaging with Deepfakes and its tools.

The Deepfakes technology was born publicly and turned into a Free Software phenomenon. It is being used, by the public, to forge pornography, create memes, do satire and perform technology demonstrations. A taxonomy of Deepfakes suggested herein also encompasses the *Deceptive Deepfake*: An imagined video so powerfully deceptive as to have serious effects on society. There are no records of such a video being made, and at this point it is certainly not easy. However, it is not entirely implausible. Nevertheless, the idea of the Deceptive Deepfake is already associated with change, no matter its actual existence. The notion of a *social imaginary* has been particularly helpful in understanding this association. A social imaginary consists of world views and social practices that reinforce each other. In the social imaginary examined herein, anyone can create or modify truths by producing Deceptive Deepfakes. This understanding of Deepfakes is essential to the publicly expressed connections between Deepfakes and the notion of *post-truth*. This association allows the post-truth condition to be reinforced by the idea of the Deceptive Deepfake, and the Deepfakes phenomenon to be reinforced by the notion of post-truth.

The post-truth condition is the source of great interest and heated debate within the field of science and technology studies (STS). Its ideas have allowed us to see any practice capable of calling forth truths to be seen as tools of knowledge production worthy of analysis in their own right. The Deepfakes phenomenon could thus be understood as the *recursive public* of the Free Software movement effectively fighting to shift the balance of power over knowledge production, while institutions of power are working to uphold the status quo.

Acknowledgements

This thesis would not have happened without the support of several people who deserve thanks:

First and most of all, thanks to my closest person in the world, Emma, for putting up with my stressed out state over the past months and still supporting me through it by any means necessary. I would undoubtedly have been twice as stressed out without you! That also goes for Hakim, who will always help me blow off some steam when necessary.

I am very grateful to my mother for always being there to support me when I need it, and to the rest of my family for keeping me in their thoughts despite my pursuing a life away from home.

Susanne, my supervisor, has helped me with valuable insights, reading suggestions and encouragements along the way, without which this thesis would not have been complete. Thank you!

My fellow students at TIK have also contributed, with lots of motivation and much needed Gurka matches between writing sessions. Thanks for keeping me sane, and let’s keep in touch.

Finally, thanks to my employer ACOS AS and Thormod for allowing me hours just flexible enough for me to be able to finish the thesis on time.

Oslo, 28th of October 2018. Tormod Dag Fikse

Table of contents

1 Introduction

Imagine for a minute that two of the world’s nuclear powers currently have unstable and unpredictable heads of state, and are currently at the peak of a brittle cold war. Picture that a state media news bulletin starts spreading online, showing one of said heads of state publicly declaring a nuclear attack on their enemy. Imagine that the other head of state is forced to respond in kind. Before the news bulletin is proven to be a fake, the two states have already engaged in mutual annihilation.

This scenario obviously has not happened, but it is a scenario that earlier this year was suggested by a university professor and spread by the media (Christian, 2018). Similar imagined scenarios involving deceptions of the public have also made the rounds online. The reason is this: In the last twelve months, high quality face-swapped videos have started circulating on the Internet, showing unexpected situations, like politicians doing odd speeches and celebrities performing in pornography. Consequently, commentators have warned that misuse of face-swapped videos potentially could have serious consequences. Whether or not the imagined nuclear apocalypse is hyperbole, the technology seems to be sparking an idea of a world in which false videos are wreaking havoc because society is not sufficiently aware that we can no longer trust what we see in a video. The videos and the surrounding phenomenon has been dubbed Deepfakes, alluding to the fact that they are fakes produced by deep learning – a form of machine learning where software algorithms learn to recognize data representations. In this context, the data representations are human faces

Such technological innovations are arguably of great interest in academics, for example to scholars who are urging for research on manipulation by communication technologies (e.g. Enli, 2015, p. 132). This thesis is written in the tradition of science and technology studies (STS), which acknowledges the profound effect technological practices and material arrangements are having on society, while at the same time rejecting mere technological determinism (Wyatt, 2007). That is, STS rejects the notion that technologies independent of society cause social change in a linear way. Instead, the tradition emphasizes an idea of co-production where technology and society are continuously producing or affecting each other in more complex ways. Instead of simply looking at technology or practices in isolation, STS is concerned with co-produced phenomena we could define as *socio-technical* practices (e.g. Jasanoff & Kim, 2015). STS is also concerned with the co-production of science and knowledge. Both are seen to be produced through socio-technical practices, and they are not simply representations of an external truth about the world. STS scholars are currently engaged in debate over the post-truth condition, where (so-called) fake news have consequences, and facts no longer matter like they presumably used to. Voters are being subjected to mediated realities that are designed to sway them toward specific candidates or referendum alternatives. An industry of voter manipulation has been exposed, represented by companies like Cambridge Analytica who have been contracted to manipulate voters in the US presidential election, the UK Brexit referendum and a long list of other elections all over the world (e.g. Cadwalladr, 2018). Similarly, reports have surfaced detailing Russian institutions—so-called troll factories—who are using the advent of social media for propaganda purposes, by producing politically loaded sentiments in the form of opinions, memes and falsified news reports (e.g. Walker, 2015).

The aim of this thesis is to understand the complex ways in which such societal practices are associated with the Deepfakes phenomenon. To this end, I will give a thick description of a technology that is oft-discussed, but also unfamiliar to many. I choose to formulate two specific research questions as such:

1. **What is the Deepfakes phenomenon, and how has it developed?**
2. **What socio-technical practices have formed around the Deepfakes phenomenon, and how are they related to the idea of post-truth?**

I have attempted to shed light on these questions by situating my empirical findings on the socio-technical practices within a framework belonging primarily to science and technology studies (STS), which I will detail in chapter 2.1. My empirical work has been done mainly through the Internet by engaging with the phenomenon through a wide variety of activities like watching videos, reading discussions, looking at source code, sifting through digital archives and experimenting with software tools. These activities make up a set of ethnographical methods that I will deliberate in chapter 2.2, describing them in terms of ethnographical theory. What follows thereafter is my ethnographical account where I have aimed to weave together empirical findings, literature and analysis into a single coherent story. I have divided the story into different chapters by topic, each rooted in an aspect of the empirical findings. The story starts at the birth of the Deepfakes phenomenon and continues to recount its various manifestations and their reception by the public. Then, it describes the technological system that Deepfakes rely on, examines the role of the Free Software community, and paints a picture of my experience with creating Deepfakes. Finally, it details my primary analysis of the phenomenon’s relation to the idea of post-truth, and how the technology could be seen as a public tool of knowledge production.

2 Framework and methods

2.1 Science and technology studies

This thesis has been written in the STS tradition and concerns itself of matters of the digital sphere. That makes the thesis a form of *digital STS*. Digital STS can be divided into two subcategories that further specify exactly how the research is digital: *first-order digital STS*—applying traditional STS methods on matters of the digital, and *second-order digital STS*—using methods and data that are inherently digital (Fikse, 2018). Methodologically, this thesis belongs primarily to *first-order digital STS*, which means that I have applied traditional social science methods on matters that manifest primarily in the digital. While a few of the methods I have used are made possible by digital tools like big data, most of the work resonates well with traditional STS.

The goal of this thesis is in part to explore how the Deepfakes phenomenon is manifesting itself through practices, and to give what anthropologists call a *thick description* of the workings of these practices. A thick description does not rely on factual description alone, but also on interpretation and analysis (Geertz, 2000, pp. 3–30). Furthermore, the thick description is based on an understanding of human culture as *semiotic* – as constituted by meaning-making through relations between signs and symbols:

> Believing, with Max Weber, that man is an animal suspended in webs of significance he himself has spun, I take culture to be those webs, and the analysis of it to be therefore not an experimental science in search of law but an interpretative one in search of meaning. It is explication I am after, construing social expression on their surface enigmatical. (Geertz, 2000, p. 5)

The cultural context, or even contexts, is an important factor in the interpretation, and Geertz points out that the meaning of an action or a symbol can vary according to the subjective cultural context or subjective web of significance in which it is understood.

I started my journey through the Deepfakes phenomenon by diving into the empirical material without a fixed idea of what theoretical work to draw on. My empirical work was generally situated within an STS framework from the start, but most of the specific theoretical reflections were made well after the empirical work was finished, as academic readings started to resonate with my findings. Similarly, I started out without a definite limit of methods to apply. With thick description as the goal, there was at the outset an uncertainty as to what forms the phenomenon would take, and what ways of research would be most applicable. By remaining flexible about what methods to employ, or at the very least choosing a flexible set of methods, I have been able to follow the leads I believe to give the most interesting answers.

2.1.1 The debates of publics and post-truth

One type of STS approach in particular has been of great important to the thesis. This is an approach to understand public opinion and democracy which stems from the heritage of ideas like the public sphere of Habermas (1991 in Kelty, 2008) and the publics of Dewey (1929) and Lippmann (1922, 1925). The so-called Lippmann-Dewey debate started when Walter Lippmann published his books in which he cast doubts on the public's ability to understand complex matters of policy and make the right choices for the good of all.

Another focal point of the thesis is the idea of post-truth, which is defined by the Oxford English Dictionary as "Denoting circumstances in which objective facts are less influential in shaping political debate or public opinion than appeals to emotion and personal belief" (OED Online, n.d.). This idea arguably touches central tenets of the STS field, and has thus sparked heated debate between scholars (Collins, Evans, & Weinel, 2017; Fuller, 2017, 2018; Jasanoff & Simmet, 2017; Lynch, 2017a; Sismondo, 2017b). Most notably, the debate arose after Steve Fuller, professor of Social Epistemology at the University of Warwick, challenged the field of STS to "embrace our responsibility for the post-truth world" and "do something unexpectedly creative with it" (Fuller, 2016). Fuller did not accept the popular pejorative description of post-truth, as he put it. What Fuller was referring to specifically here was the Oxford English Dictionary's definition, which is much in line with STS scholar Sismondo's five "common post-truth tropes":

1. The emotional resonances and feelings generated by statements are coming to matter more than their factual basis.
2. Opinions, especially if they match what people already want to believe, are coming to matter more than facts.

3. Public figures can make statements disconnected from facts, without fear that rebuttals will have any consequences. Significant segments of the public display an inability to distinguish fact and fiction.
4. Bullshit, casual dishonesty and demagoguery are increasingly accepted parts of political and public life; this should not, however, be confused with ordinary lying, which is nothing new.
5. There has been a loss of power and trust in traditional media, leading to more fake news, news bubbles and do-it-yourself investigations.

(Sismondo, 2017b)

However, Sismondo's five tropes were in stark contrast to Fuller's view of the post-truth condition. Fuller's understanding was that post-truth is nothing new, but simply the ebb and flow of an ongoing power game between 'lions' (those who wish to uphold the status quo) and 'foxes' (those who want to change the rules of the power game in their favor). Fuller's point was basically that dominant ideas in STS, particularly the principle of symmetry, have laid the groundwork for the idea that knowledge production or valid knowledge is not reserved for those with institutional authority. Ever since Kuhn introduced the idea of paradigms in science and knowledge production, it has been widely held that knowledge is relative to the current world view and scientific framework (Kuhn, 1962, 1970). The principle of symmetry furthermore holds that the truth or falsity of a belief should be judged in the same terms, using the same methodology. This in turn effectively acknowledges that there are social reasons that truths are held to be true. In Fuller's view, the state of post-truth has followed naturally from an increased acceptance for social construction of knowledge—and from an increase in the public's ability to produce knowledge:

> My own view has always been that a post-truth world is the inevitable outcome of greater epistemic democracy. In other words, once the instruments of knowledge production are made generally available—and they have been shown to work—they will end up working for anyone with access to them. This in turn will remove the relatively esoteric and hierarchical basis on which knowledge has traditionally acted as a force for stability and often domination. (Fuller, 2016)

Fuller notes that that a state of post-truth will reduce the dominating yet stabilizing effects that a world of truths has had, for better or worse. Such a change is nevertheless welcome to

Fuller, who seems to want to move toward the establishment of a new order based on a knowledge production of the people, as opposed to a knowledge production of the experts. In Kuhnian terms, this would let the public affect, or even define the paradigm of knowledge production.

Fuller's field of social epistemology touches at the heart of the idea of publics as it is expressed through the Dewey-Lippmann debate. As Steve Fuller points out, Lippmann's doubt of the public was amplified by the prospect of new audiovisual media technologies:

> It is worth recalling that in the middle third of the twentieth century, such promoters of an expert-steered mass society as Walter Lippmann (1922) and Alfred Schutz (1946) were concerned that the increasingly sensory character of the news media – whereby print is supplemented if not replaced by sound and vision – would result in people acquiring a kind of 'pseudo-experience' that would lead them to think that they know more than they really do, simply because they can claim to have 'seen' or 'heard' certain things broadcast in the media, which is then mixed with genuine personal experience to generate what they regard as politically valid judgements. (Fuller, 2018, p. 21)

The notion that media can give us the wrong idea about reality was certainly not reservedly a 1920s one. These old ideas arguably resonate well with today's industrialized deception (chapter 1). Today, it is widely held that "although we base most of our knowledge about our society and the world in which we live on mediated representations of reality, we remain well aware that the media are constructed, manipulated, and even faked" (Enli, 2015, p. 1). Lippmann was also aware of this at the time, and as Fuller points out, Lippmann's view was that we would therefore be better off leaving the choices with expert leaders suited to the task. This brings us to John Dewey (1929), who countered with the argument that experts were no guarantee for rationality or against abuse or self-interest. Dewey and Lippmann had similar opinions on what should be the practical role of the public, namely to elect officials, but Dewey had a more marked emphasis on the necessity of an informed public balancing out the power of elected officials. Today, Fuller has exceeded Dewey's emphasis, and argued that the post-truth condition not only follows naturally from STS, but also that it is a welcome change toward a democracy of knowledge production.

Some STS scholars have chosen to deny Fuller's connection between STS and post-truth completely, like Sismondo (2017a) who accused Fuller of missing a central tension in STS, namely that democratizing knowledge production doesn't have to be easy, and should still be associated with relevant "infrastructures, efforts, ingenuity and validation structures". Sismondo predicted "If the post-truth era starts by blowing up current knowledge structures, then it isn't very likely to be democratization, and in fact most likely leads to authoritarianism." Sismondo acknowledged that a state of post-truth is possible, that the danger might be real, not least due to voter manipulation by existing authorities, but also because of strong anti-expertist sentiments. Fuller has received support by other STS scholars, like Collins, Evans, & Weinel:

> To claim that STS never came down on the side of politics rather than technical expertise is, itself, to try to do some serious political work. If we want to avoid being accused of falsifying our own history we have to admit that for much of the time the views STS was espousing were consistent with post-truth irrespective of their authors' intentions or their causal impact. The flaw in Sismondo's analysis is the idea that post-truth is 'easy' and that this is what separates its crude politics from the more sophisticated analysis of STS. But post-truth is hard work: look at the work Trump and his supporters are putting into it beyond simply working a Twitter account; look at the work Joseph Goebbels did to tell 'the big lie'; look at the work that had to be imagined to organize George Orwell's 'Ministry of Truth'. (Collins et al., 2017)

Collins, Evans & Weinel point out that doing post-truth is hard work—it is not just simple lies, but sophisticated knowledge production work to be analyzed in its own right, whether one likes it or not. However, they posit that STS should take the role of understanding legitimacy, and effectively judge if it is being used correctly, for example in cases where the would-be experts have the wrong idea:

> In all these examples, understanding who can legitimately contribute to expert debate requires social scientists to use their special understanding of the formation of knowledge to reject the misuse of expertise by certain elite experts and give credit to the work of low status, experience-based experts. (Collins et al., 2017)

In essence, they argue that discarding all current practices of expertise in favor of (a misunderstood) democracy would lead to our way of life being destroyed, even if we should be working to understand why *some* practices of expertise are not optimal.

For the later analysis it might be useful to distinguish between facts and truths as such: The legitimacy of a fact comes from its *scientific* verifiability, while the legitimacy of a truth comes primarily from *social* authority that may or may not rely on facts (Sørensen, 2017). Thus, post-truth should perhaps instead be called post-fact, due to the shift in emphasis from facts to truths. Other terms have also been suggested, for example by Lynch (2017b), who describes the situation as a state of alternative *alt-truths*, spurred by *alt-facts*.

2.1.2 Free Software and social imaginaries

Kelty's ideas on Free Software have also contributed to this thesis. The terms Open Source and Free Software are often used interchangeably, with mostly subtle differences of emphasis. Both describe software made public; software that can more or less freely be inspected, edited and made use of without significant restrictions. Both terms are also in opposition to closed source, which designates software for which the source code has been obfuscated and where the distribution, costs and changes to the software are controlled by someone. In addition, the two terms carry with them a set of characteristics. Kelty (2008) has suggested five core practices that I will sum up as follows, describing them in my own words:

- **Sharing source code:** Allowing people to obtain a copy of the source code, enabling them to see exactly how a software has been made, and use it for themselves.

- **Conceptualizing open systems:** Providing a systemized way of free adaptation and modification, for example by way of separating the code into different forks that can be developed in different directions.

- **Writing licenses:** Specifying rules for the use, adaptation and modification of the source code through a formal license document.

- **Coordinating collaborations:** Inviting people to contribute, and organizing it through source control software that control the merging of code and allow the tracking and discussion of development issues.

- **Fomenting movements:** self-reflection on the nature of the practices and their implications.

Kelty is careful about freely designating Free Software as a movement, preferring instead to describe it as a *recursive public*. He explains this term—in the words of Taylor—as a type of social imaginary. A social imaginary is a parallel to the idea of publics, but with an extra emphasis on merging material practices and ideas together. When these two come together, a social imaginary forms, where "one cannot distinguish the two in order to ask the question Which causes which?" (Taylor in Kelty, 2008, p. 39). Kelty's specific social imaginary describes the recursive public of Free Software as one "concerned with the maintenance and expansion of the infrastructures that allow them to come into being in the first place" (2008: 17). Free Software exists because of infrastructure such as the Internet and a wide array of tools and practices that the recursive public by its own idea can and should use in order to strengthen its own importance and defend its own practices. This is a form of co-production where ideas and material practices—like those that surround a technology—reinforce each other. In other words, the idea that Free Software practices gives the power of knowledge to its public is reinforcing the practices themselves, and vice versa. This concern for expanding Free Software infrastructures is deeply entwined in geek culture, and Kelty also claims that the recursive public of Free Software has changed the control over the creation and dissemination of knowledge. This effectively reorients the balance of power between governing bodies and the recursive public.

2.2 Ethnography

At the outset of this study, it seemed it could be possible to shed light on the Deepfakes phenomenon in a number of different ways: by analyzing texts or videos, by following discussions on- or offline, by performing interviews, by observing, by tracing the circulation of videos, by examining log files, by partaking in the creation of videos and so on. It seemed likely that the best way forward would be combine several of these methods. This flexibility is a key trait of ethnography:

> "The differences from other approaches relate as much to the lack of advance specification of method as they do to the actual methods used. By refusing to decide in advance what will be most interesting to explore in the setting, the ethnographer

> remains open to novel discoveries about the unique ways that a particular way of life might be organized and to the prospect that activities may make sense in surprising ways.” (Hine, 2015, p. 25)

In other words, not deciding all matters of methodology beforehand is a valid approach, and even possibly a healthy one. Ethnography is not clearly defined, and if it is defined the focus is on open-ended and exploratory deliberations on a specific practice or set of practices (Hammersley & Atkinson, 1986, pp. 2–3).

The chosen methods would either way not give a complete account of the Deepfakes phenomenon. Not only would a complete account be impossible given the time constraints for the thesis, there is also not just one experience of any online phenomena, or indeed any phenomena. It is experience as part of Geertz’ web of significance. Not being able to describe a single reality is part of the uncertainty an ethnographer has to live with (Hine, 2015, p. 88). I have not ended up with a singular truth about Deepfakes, as I have not aimed to. To the contrary, a thick description takes subjective interpretation into account. At the same time, ethnography is “very well suited to giving us a critical stance on over-generalized assumptions about the impact of new technologies” (Hine, 2015, p. 2). While it might be premature to describe assumptions about Deepfakes as over-generalized, my methods have certainly helped me avoid the pitfall of generalized conclusions on the impact of Deepfakes. The flexible nature of this ethnographic approach has allowed the findings to decide whether to confirm popular assumptions or provide a different or more complex point of view.

The goal has been for the thesis to utilize ethnography’s ability to add to the understanding of the phenomenon by giving a “theoretically enriched form of description” (Hine, 2015, p. 59). Hopefully, the reader will agree that this has not only interpreted the phenomenon as such, but also added to theoretical discussions.

While the exact choice of and emphasis on methods was decided during the exploration, it was certainly possible to give an account in advance of certain aspects of ethnography that should be covered. The following topics are methodological corner stones of ethnography that I have chosen to employ in my study, in some form or other.

2.2.1 Participation

One important way of gaining knowledge of practices through ethnography is for the ethnographer himself to participate in the practices. This is because it will give the ethnographer the possibility of observing in detail what the practices are about, and learning-by-doing, instead of relying on the participants own accounts (Hine, 2015, p. 55, 73). This is done by taking part in the same activities as the people involved in the practices that are under scrutiny, to allow the ethnographer to be immersed, share the feelings of the involved people and test his interpretations continually (Hine, 2015, p. 19). My participation has been two-fold: Firstly, I have made use of the technology itself, experimented with the creation of my own videos, taken part in the communities dedicated to the technology and learned first-hand what it means to create a Deepfake. Secondly, I have actively participate in deliberation on the impact of the technology, seeking out opinions and discussing the topic with my peers to develop my understanding of the phenomenon. The chosen form of participation has made my study partly autoethnographical. My own participation has made me a part of the phenomenon I have been studying, and my own personal experiences has been an important part of the analysis.

In order to participate, one would traditionally begin by asking: “where do I go?” (Beaulieu, 2010), which brings us to the field site.

2.2.2 Field sites

In traditional ethnography, the field site is the space in which the studied social practices occur (Burrell, 2009, p. 182). Traditionally, ethnographers would live or hang out in the field sites over long periods of time to attain the necessary degree of participation and immersion. The choice of field site is an important one, but the chosen space is no longer necessarily or simply a geographical one. Instead it is increasingly thought of as an artfully constructed network of locations, physical or otherwise, where the objects and subjects of study manifest (Burrell, 2009, p. 192; Hine, 2015, p. 60). Particularly in studies concerning matters of the Internet, the field site is not easily defined. Studies must take on a slightly different form when the subject matter exists largely as a network or an online phenomenon, and not in a geographical location. When the social practices are partially or primarily mediated, the connections transcend the online/offline divide (if indeed such a divide even exists). The ethnographer's aim then is to immerse herself in whatever activities are relevant, through whatever media is relevant, to get as close as possible to the experience of the subjects (Hine, 2015, p. 56). In this form of ethnography, the field site is thus defined by a form of *co-presence* instead of physical *co-location* (Beaulieu, 2010). Allowing an ethnography of co-presence instead of co-location is essential to this thesis:

> […] co-presence, as a way of deploying ethnography, can enrich our understanding of key concerns for STS: what counts as knowledge, and why, because this approach to fieldwork may also renew interest in the question of what one is studying in STS. If the ethnographer enters into what is recognized as a space of science, then he or she can (and often does) appeal to the setting to claim that at some point, the activities encountered will constitute science (Beaulieu, 2010).

When we allow ourselves to look for knowledge production in mediated settings where we previously have not, we also open up for the possibility of considering its practices knowledge production. Looking at websites has not always been considered proper ethnographic fieldwork. However, STS scholars are increasingly taking mediated settings seriously by moving from a spatial focus to instead focusing on a stream of practices (Beaulieu, 2010), wherever they exist. This means that social practices of knowledge production may manifest in a database or algorithm as well as in physical space (Star, 1999). Consequently, it also means that participation can happen without being present at the same time. Mediated experiences can be time-shifted—they can have different temporal forms. They may also largely rely on infrastructures, technologies and existing practices (Beaulieu, 2010). In this thesis some of the prerequisites of Deepfakes will be considered briefly, but a full analysis of the elements sustaining the interactions is well beyond the scope of a single study.

2.2.3 Field notes

Ethnographic studies are traditionally documented continuously by way of field notes, as has also been the case for this thesis. However, there are important differences between studying social exchanges in person and studying (traces of) social activity on the Internet. This has forced me to adapt the traditional way of writing field notes. Specifically, the choice of strategy for taking field notes is influenced by the over-abundance of data available online.

The Deepfakes Club forum—which is one of many sites informing my research—alone has more than 350 different discussion threads and 1800 posts at the time of writing. It would be an extremely daunting task to partake in all these threads and for each one write up rich, detailed notes such as the sketches or episodes often used in traditional ethnography (Emerson, Fretz, & Shaw, 2011, pp. 75–79). Such a strategy would not only limit the possible scope of inquiry, but it also seems quite unnecessary given that the data in many cases is right there, ready to be viewed again at any time. Of course, there is always the risk that the observed material may be removed or become inaccessible at some point. A practical compromise for this study has been to make notes on the general analysis of the more mundane exchanges, while picking out specific interesting or exemplary cases to be described in more detail or saved for later review. This strategy has demanded that I be extra careful about the balance between efficient abstractions and premature analysis. If cases had been analyzed too quickly, and details left out too often, the result might not be the open-mindedly compiled thick description that I wanted to end up with.

2.2.4 Chosen setting

My findings and analysis have allowed me to define my setting as a social imaginary and my subjects as a manifestation of the recursive public, but this was not possible at the outset. Only after diving into the Deepfakes phenomenon and exploring it has it been possible to determine the nature of my setting. My field site has not been a place. Deepfakes are everywhere and nowhere. Talk of Deepfakes happens across sites and platforms, videos are cross-posted and ideas are coming into being in parallel in a myriad ways. My way into the phenomenon has been of limited importance, while following the network around has been absolutely essential.

My study has taken place wherever I might easily find traces of the socio-technical practices surrounding Deepfakes and those who enable their production, produce them, consume them and discuss them. Certain kinds of people have proven essential informants to this thesis by leaving traces of their practices. These are particularly the technologically curious experimenting with the development and diffusion of Deepfakes, the journalists and academics commenting on the phenomenon, and the legislators reacting to it. I have assumed that the socio-technical practices most closely related to the technology are mediated practices. The development of the algorithms and the video experimentation has not taken place in a co-locational social setting. People don't generally gather together in a room to work on Deepfakes together. Instead, the communities and collaborations have formed online. Thus, certain specific parts of the Internet have proved to be of particular importance to gathering empirical material on the Deepfakes phenomenon. These are a wide range of discussion forums like Deepfakes Club, code sharing services like GitHub and video sharing sites like YouTube. In addition, some traces have been found in Internet archives like archive.is. A long list of digital news publications, as well as paper-based ones, have also proved useful—like those of The Guardian. Furthermore, much learning has come from interacting directly with the Deepfakes technology and its tools in my own Deepfakes lab.

2.2.5 Storytelling ethics

Ethnography is, by its very nature, a style of research in which the researcher is prominent. The choices as to what findings to focus on and what leads to pursue, are actively shaping the outcome. While my ideas have changed as the empirical findings have surfaced, I am sure that I have followed threads that have seemed promising in reinforcing my initial ideas. This means the final product of the ethnography is a form of story, told by myself as the researcher. Ethnography does not claim to create a neutral account of the studied culture. This however, does not necessarily make it any less worthy than any other form of research. Instead, as Hine (2015, p. 20) remarks, its authenticity stems from the ethnographer's direct experience through immersion. In line with this, my thesis describes my process of knowledge-generation in a personal and in-depth way, expressing the involved processes of thought thoroughly. As Emerson et al. (2011, pp. 245–246) have suggested, I have allowed myself to be visible in my field notes and the thesis.

Secondly, I am open about how my reworking of the field notes are a reconstruction based on choices of what to emphasize. While fictitious accounts can sometimes accompany an ethnography, there is no such material in this thesis, with the exception of the imagined situation laid out in the introduction. It has been important for me to retain the integrity of my subjects when representing their statements and opinions, and an important starting point has been to be aware of the significance of my own interpretation. My own experiences, views and priorities has inevitably shaped my finished ethnography (Emerson et al., 2011, p. 247), just as I assume my field of study (STS) has affected my analysis and personal opinions. Therefore, in the name of transparency, I have allowed my own voice to be an important part of the story. This is in line with the conventions of ethnographic study. Hopefully this has been done without becoming self-indulgent (Hine, 2015, p. 20).

Lastly, to reduce the risk of disturbing the privacy of my subjects, I have decided to anonymize all but a select few of the most public informants. This means that I have left out names or usernames where they are not important to the story. More importantly it also means that I have made sure that most statements cannot be retraced to their source by use of search engines, thereby possibly identifying the informants. To this end I have paraphrased statements and left out direct references to them, instead referencing my own field notes. The Norwegian Centre for Research Data has been notified of my research project and the extent of personal data processing involved. I have also chosen to leave little trace of my own participation in the (non-spatial) field, to limit the risk of affecting the phenomenon through my participation.

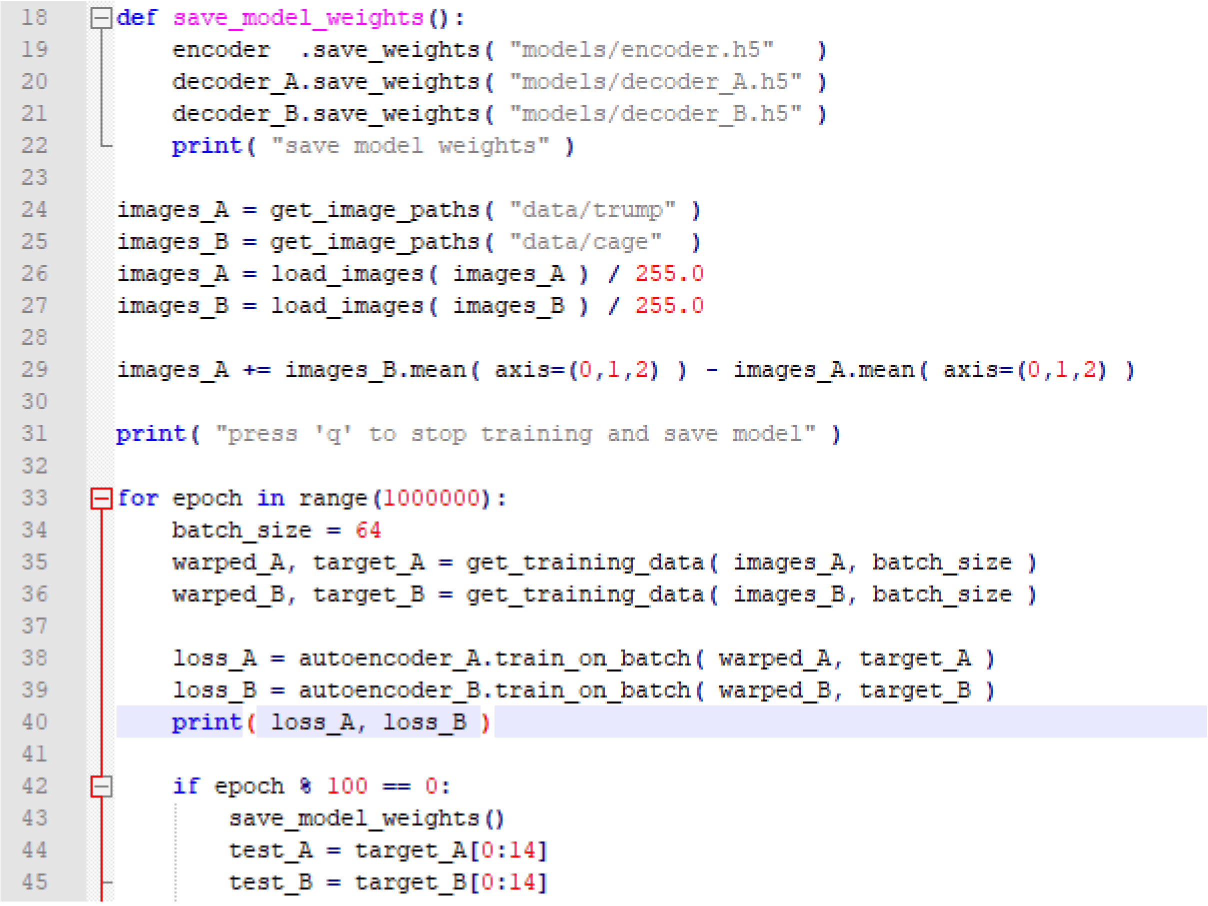
3 The story

3.1 The practices

3.1.1 The birth: Deepfakes coming to life

The start of the Deepfakes technology is usually attributed to an anonymous user on the social media platform Reddit. Reddit is a platform comprised of more than a million smaller *subreddits* dedicated to different topics. These are small communities of content sharing and discussion—one of which was created by the user only known as u/deepfakes in November 2017. The community was aptly named r/deepfakes, and this is where the first faceswaps using the Deepfakes algorithm started circulating. This is also where the first rendition of the Deepfakes algorithm—the software behind the videos—was made public. This subreddit has since been removed by Reddit leadership (Hawkins, 2018) after the site rules were updated to ban what they called involuntary pornography (chapter 3.1.2). As it happens, I was a frequent visitor of Reddit at the time, and witnessed parts of the birth first-hand before deciding to write this thesis, since some of it reached the Reddit front page and was also exposed to those of us who didn't follow r/deepfakes. Instead of relying on my own memory or media reports, I wanted to return and explore this moment that represents the public birth of Deepfakes, and look for reasons the person or people behind Deepfakes had for making the technology public. Given the removal of the subreddit, this was no easy task. However, by searching several different web archiving services, most of them unsuccessfully, I have managed to find some snapshots of pages belonging to r/deepfakes through the web archive archive.is. In these snapshots, the moment when user u/deepfakes first shared his code is preserved:

The archive shows that u/deepfakes uploaded a package through an anonymous file sharing service, and posted it on Reddit for the community members to start experimenting with (Field notes, July 11 2018). The link to the shared file was still active at the time of writing, and I was able to download the package, which contained a set of example data based on images of Donald Trump and Nicholas Cage, as well as 398 lines of code written in Python (Figure 1), of which 170 lines were licensed from other coding projects.



```
def save_model_weights():
    encoder  .save_weights( "models/encoder.h5"   )
    decoder_A.save_weights( "models/decoder_A.h5" )
    decoder_B.save_weights( "models/decoder_B.h5" )
    print( "save model weights" )

images_A = get_image_paths( "data/trump" )
images_B = get_image_paths( "data/cage"  )
images_A = load_images( images_A ) / 255.0
images_B = load_images( images_B ) / 255.0

images_A += images_B.mean( axis=(0,1,2) ) - images_A.mean( axis=(0,1,2) )

print( "press 'q' to stop training and save model" )

for epoch in range(1000000):
    batch_size = 64
    warped_A, target_A = get_training_data( images_A, batch_size )
    warped_B, target_B = get_training_data( images_B, batch_size )

    loss_A = autoencoder_A.train_on_batch( warped_A, target_A )
    loss_B = autoencoder_B.train_on_batch( warped_B, target_B )
    print( loss_A, loss_B )

    if epoch % 100 == 0:
        save_model_weights()
        test_A = target_A[0:14]
        test_B = target_B[0:14]
```

Figure 1: Excerpt from original Deepfakes model training code shared with the Reddit community in November 2017.

In the Reddit post sharing the code, the author(s) didn't make their underlying motives entirely clear, but they did share some reasoning:

> u/deepfakes:
> As you can see, the code is embarrassingly simple. I don't think it's worth the trouble to keep it secret from everyone. I believe the community are smart enough to finish the rest of the owl [sic]. (Field notes, July 11 2018)

The author did not seem to think much of his own work, and implied that it should be developed further by the subreddit community. He did not stay an active part of this development, but before going silent he chimed in on a discussion about consequences. Another Reddit user had foretold that the pornographic celebrity fakes would eventually lead to people faking pornographic videos of colleagues or classmates. In their idea, it would lead to rampant sexual harassment, and women in particular would have to refrain from being photographed or uploading photos to social media. The author replied:

> u/deepfakes:
>
> I think it has already happened. The only difference is that it's me or some one else to put the last straw that breaks the camel.
> It's not only me but a whole industry studies on how to generate realistic images. In one or two years, half of Photoshop's tools will be based on machine learning. Animators will have machine learning based tool to create natural character animations. People will train model to detect fake images, and other people will train model to create undetectable fakes. Face swapping is nothing compare to creating realistic 3D avatars and putting them in virtual reality. (Field notes, July 11 2018)

The author claimed his contribution was of little importance in causing the unwanted applications of machine learning, because so many other related technologies are being researched by the industry. Indeed, research on a technology with quite similar applications as Deepfakes, called Face2Face, was published a year prior to Deepfakes (Thies, Zollhöfer, Stamminger, Theobalt, & Nießner, 2016). Other machine learning technologies similar to Deepfakes have been developed, and continue to be developed in parallel today (Mukhopadhyay, Shirvanian, & Saxena, 2015; Pham, Wang, & Pavlovic, 2018; Vougioukas, Petridis, & Pantic, 2018).

Before going deeper into the phenomenon it might be helpful to summarize a short glossary of the most frequent uses of the term Deepfakes, to avoid confusion in the following parts:

- **Deepfakes**: Face-swapped videos made using the Deepfakes technology
- **Deepfakes technology**: the set of tools required to create Deepfakes
- **r/deepfakes**: The first Reddit community (subreddit) dedicated to the technology
- **u/deepfakes**: The Reddit user sharing the original source code that the Deepfakes technology is based on
- **Deepfakes phenomenon**: The entire set of socio-technical practices surround the Deepfakes technology and videos

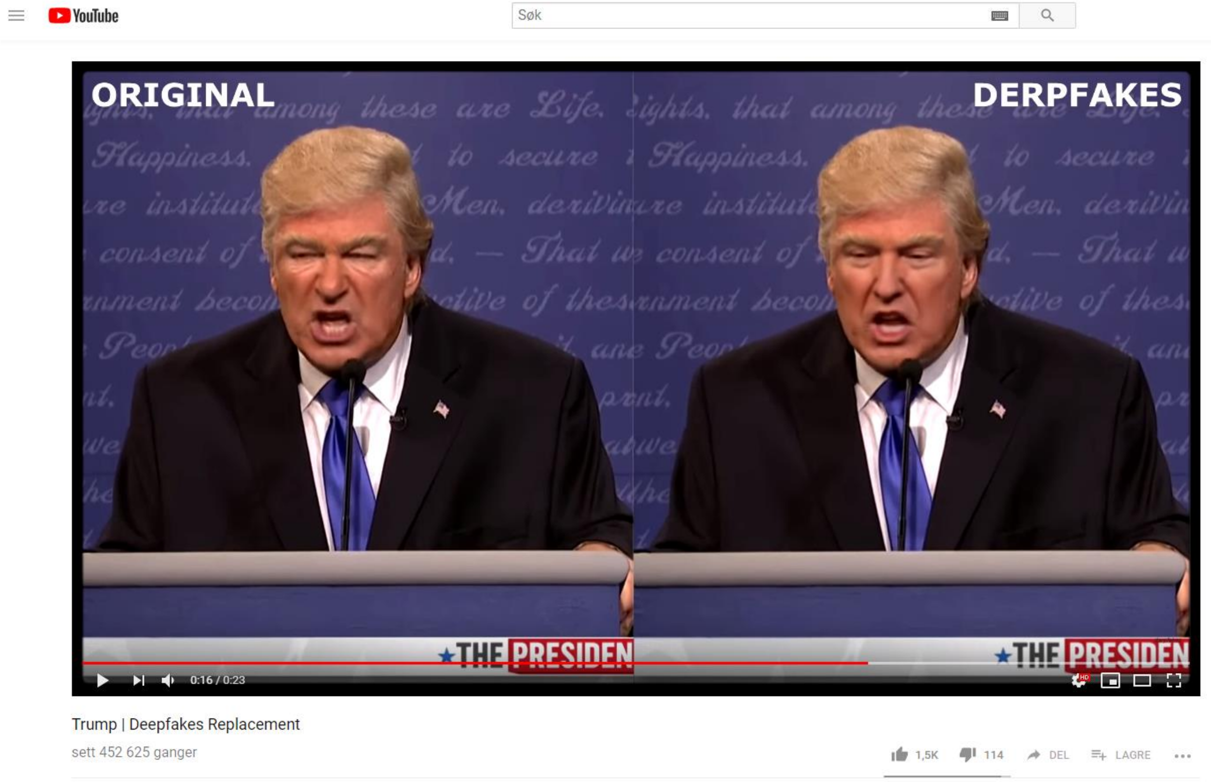
3.1.2 The videos: A taxonomy of Deepfakes in the wild

An essential part of the empirical work I have done is to trawl the web for examples of Deepfakes. Through this exploration I have noticed a pattern of identifying video traits frequently occurring together. This pattern has led me to suggest a simple taxonomy of my findings. The taxonomy is not meant to be a final overview of possible Deepfakes, but mainly meant as a tool to point out the most prominent differences in various ways of using the technology. It will show useful in later analysis of the phenomenon at large. An important point that I will highlight in each category is the level of fakery involved, using terms from Gunn Enli's theory of mediated authenticity (2015).

My suggested taxonomy consists of the following categories, in no specific order:

- The Technology Demonstration Deepfake
- The Satirical Deepfake
- The Meme Deepfake
- The Pornographic Deepfake
- The Deceptive Deepfake

Technology Demonstration Deepfakes are examples made to demonstrate how the technology works. These often feature side-by-side comparisons of the original video and the faked video. One such example is made by YouTube user derpfakes and shows actor Alec Baldwin doing a Donald Trump parody on SNL, compared to the same video with Trump's face swapped out for Baldwin's (Figure 2). Some of these are explicitly made to point out the faked parts from the start, to show exactly what the technology has done to the original video. Others may contain an *authenticity puzzle* (Enli, 2015, p. 18), which invites the audience to identify the authentic parts from the fake parts, usually before the puzzle is solved and the fakery exposed in the video's conclusion.



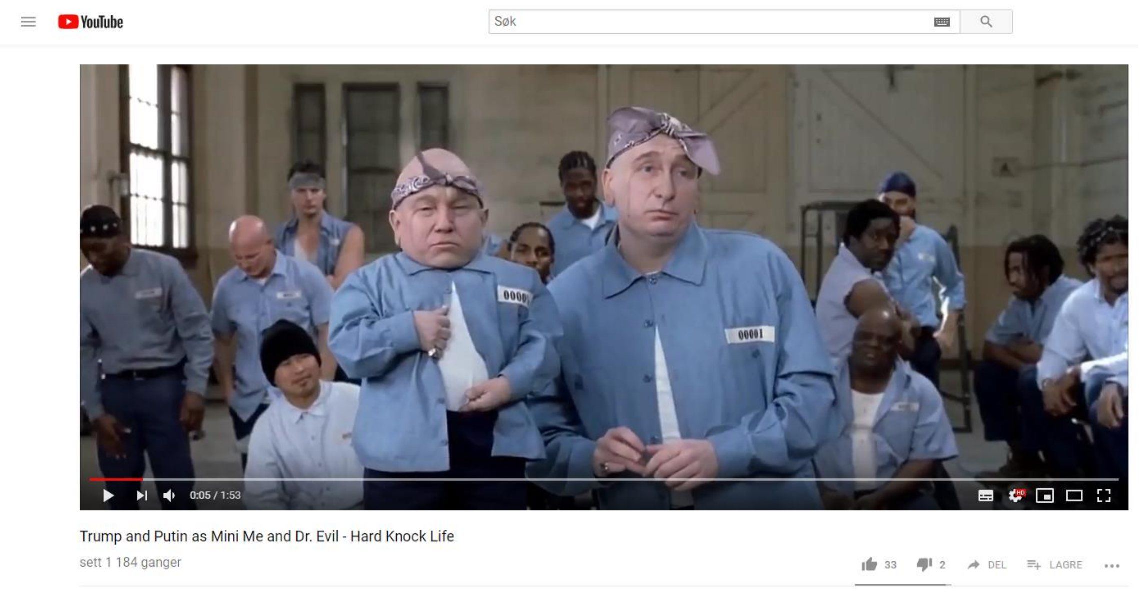
Trump | Deepfakes Replacement

sett 452 625 ganger

Figure 2: Screenshot of technology demonstration Deepfake featuring Alec Baldwin as Donald Trump—and Donald Trump as Alec Baldwin

Satirical Deepfakes are humorous or mocking videos where faces have been swapped out as a form of political or social commentary. These are not made to be deceptive, but instead rely on a form of implicit *authenticity contract* (Enli, 2015). The authenticity contract is a common understanding between producer and viewer, in this case that the video has been manipulated. The contract is based on existing genre conventions and norms, and relies on the audience's media literacy and interpretative skills to understand the rules and avoid an *authenticity scandal* (Enli, 2015, p. 18) in which audiences are deceived, resulting in problems for the audience or society. While there could be a risk of an authenticity scandal, it is not intentional.

One example of the Satirical Deepfake depicts Vladimir Putin as Dr. Evil and Trump as Mini-Me in a scene from the comedy film "Goldmember" (Figure 3). The film portrays the over-the-top caricature supervillain Dr. Evil and his miniature clone Mini-Me. This Deepfake can be understood as a satirical suggestion that Vladimir Putin is a supervillain on a geopolitical scale and that he has created for himself a less important lackey and miniature supervillain, Donald Trump. The video is referencing allegations of Russian meddling in the American elections leading up to Donald Trump's victory, as well as an idea of their shared corruption.



Trump and Putin as Mini Me and Dr. Evil - Hard Knock Life

sett 1 184 ganger

Figure 3: Screenshot of satirical Deepfake featuring Vladimir Putin as Dr. Evil and Donald Trump as Mini-Me in a scene from 'Goldmember'.

Meme Deepfakes are videos where the faces of one or more persons have been replaced as a replication or transformation of an existing, shared cultural idea. These are often simply humorous ideas, but they could potentially also act as a form of political participation (Shifman, 2014). Like the Satirical Deepfake, these are not intended for deception. Actor Nicholas Cage has been a common target of Meme Deepfakes (Figure 4), as he has long been a common target of other meme imagery. The fact that example data in the very first release of the Deepfake source code contained a set of images of Nicholas Cage (Figure 1) might also partially explain the proliferation of these Deepfakes.

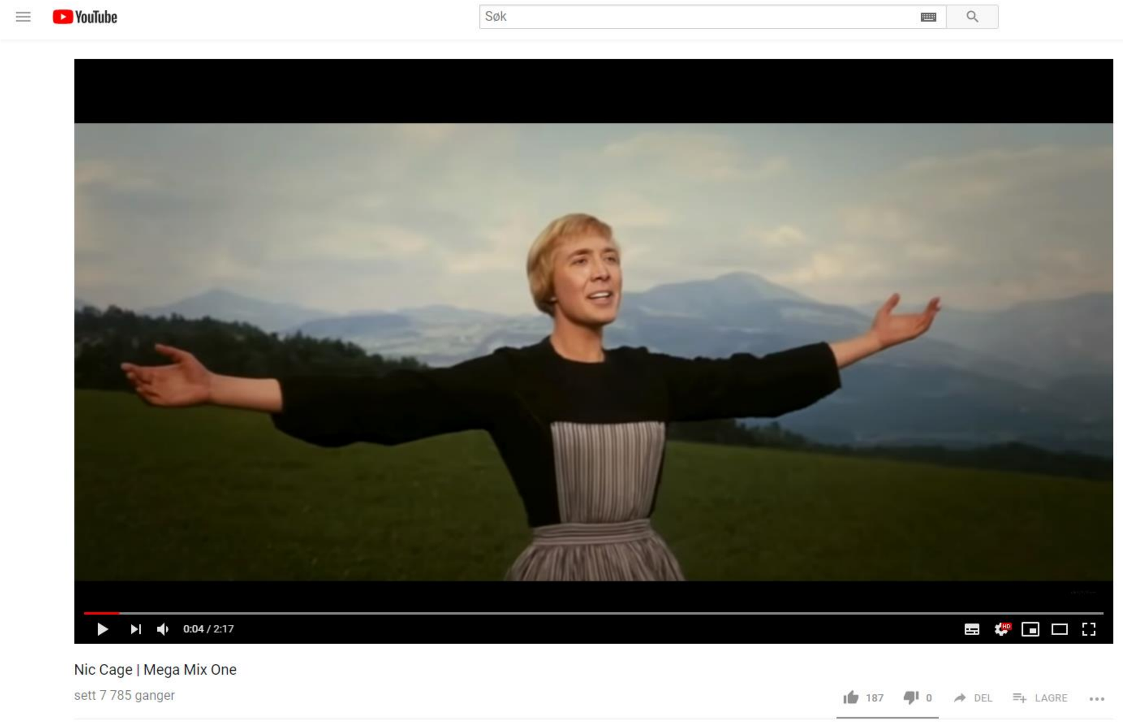
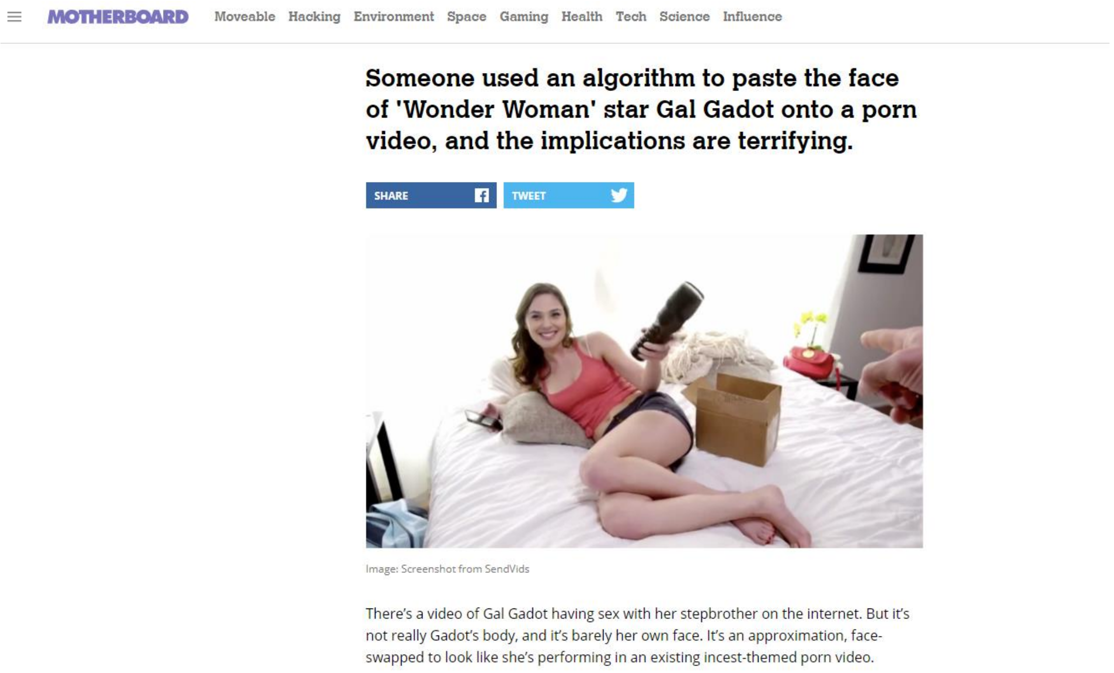


Figure 4: A screenshot of a Deepfake featuring Nicholas Cage in a variety of movies he has never played in

Pornographic Deepfakes often carry the face of Hollywood actresses (figure 3) or other celebrities on the body of pornographic actors (Figure 5). These exist in large numbers on the web. Much of the early controversy around Deepfakes focused on the lack of consent from the people swapped into a pornographic Deepfake, and the potential for the technology to be used to create revenge pornography. Such uses raise their own set of implications and questions that are out of the scope of this thesis, but that others might want to research. These Pornographic Deepfakes will often be explicitly described as forgery through their context of being shared on a site or in a category dedicated to Deepfakes. At the same time, they are likely intended to be viewed with a suspension of disbelief (Enli, 2015, p. 18). While these videos might pose as authentic sex tapes, actively trying to deceive and create an authenticity scandal, I have not focused on trawling the web to look for actively deceptive Pornographic Deepfakes. Neither have I come across any traces of them. If I did, I would rather have placed them in the next and final category.



Someone used an algorithm to paste the face of 'Wonder Woman' star Gal Gadot onto a porn video, and the implications are terrifying.

Image: Screenshot from SendVids

There's a video of Gal Gadot having sex with her stepbrother on the internet. But it's not really Gadot's body, and it's barely her own face. It's an approximation, face-swapped to look like she's performing in an existing incest-themed porn video.

Figure 5: A "safe-for-work" referencing of Deepfake video depicting actor Gal Gadot's face in pornography.

Deepfakes come in many varieties, and some are certainly hybrids of the species I've laid out so far. However, my findings, and thus the taxonomy, has a glaring lack of one specific type of Deepfake that one might have expected to find. In all the examples from my study, the forgery has been open and apparent, primarily through the context of the video or its description. However, from the social commentary on the Deepfakes phenomenon (chapter 3.1.3) I suggest to add yet another category to the taxonomy as such:

Deceptive Deepfakes is a term I suggest to describe the idea of videos made to fool its viewer into believing a forged situation involving an actor of importance to the viewer. These videos would be made with the intention of creating an authenticity scandal. They would have a political, legal or other social effect on the viewer or the actors involved, for example by faking a video of an authority figure, creating an illusion of some sort of video evidence.

My study has no findings showing the actual existence of Deceptive Deepfakes. Yet, the notion of their existence seems important enough to allow a guest appearance in the taxonomy. Most available public writings and discussions on the Deepfakes phenomenon assume that it is only a matter of time before the Deceptive Deepfakes end up changing the world. To assess whether or not that is going to happen is outside of the scope of this thesis. Nevertheless, I would like to suggest that the *idea of* Deceptive Deepfakes is already associated with change, and that has to do with how it was picked up by the media and discussed in the public. I will detail this in the following chapter, and illustrate exactly why this category deserves a place in the taxonomy.

3.1.3 The reactions: Deepfakes as a scary tool in the employ of evil

When Deepfakes videos started circulating in November of 2017, it did not take long before it was picked up by news outlets. One of the first stories on the phenomenon, if not the first, was published in the US science and technology publication Motherboard (Cole, 2017). The multimedia magazine—which aims to produce stories on the wonderful and terrifying futures of science and technology—alleged that “AI-Assisted Fake Porn Is Here and We’re All Fucked”. According to the author, “we're on the verge of living in a world where it's trivially easy to fabricate believable videos of people doing and saying things they never did”. The author cited an AI researcher on the technology having “huge ethical implications”, and had talked to pornography performers emphasizing the lack of consent from those faked into a pornographic scene. Different variants of this story soon followed by other publications all over the world.

Over time, however, the stories strengthened their emphasis on the potential political implications of the technology. In June of 2018 publications cited AI researchers at US universities who participated in wagers on whether or not the world would see its first political Deepfakes scandal in the US midterm elections of 2018 (Hsu, 2018). One such researcher, Michael Horowitz, spoke of the probability in the following way:

> Would you be that surprised if the week before the midterm elections, a Deepfake video came out that was some kind of Russian agitation designed to inflame Americans regardless of which side they were on […] In some ways, at this point I would be surprised if the Russians didn't try—the question is how much pickup it gets.

Over the summer of 2018, this view seems to have gained traction in the US government. One member of the US congress warned about Deepfakes disinformation campaigns in a speech to a think tank in July (Johnson, 2018), while two other senators addressed the issue in a hearing of Facebook and Twitter officials in September (Keane, 2018). According to senator James Lankford, "Americans can typically trust what they see and suddenly—in video—they can no longer trust what they see because of the opportunity to be able to create video that's entirely different to anything in reality has now actually come". The US Defense Department's DARPA have also followed suit through their Media Forensics program, calling for the creation of techniques for detecting video manipulation by machine learning. Such techniques have already been made (Knight, 2018; Li, Chang, & Lyu, 2018).

As of the time of writing, the reporting on the Deepfakes phenomenon has almost consistently included either the term *post-truth,* or the term for its popular media vessel, *fake news*. Most media attention has speculated that Deepfakes would at some point make it impossible to discern between real or fake videos. This would in turn, if we are to believe the rather consistent public narrative, be a danger to our democracy. The implicitly assumed trustworthiness of a video in, say, 2015, would no longer hold. This view was well detailed in a thorough blog post by two US professors of Law:

> The spread of deep fakes will threaten to erode the trust necessary for democracy to function effectively, for two reasons. First, and most obviously, the marketplace of ideas will be injected with a particularly-dangerous form of falsehood. Second, and more subtly, the public may become more willing to disbelieve true but uncomfortable facts. (Chesney & Citron, 2018)

The issue at hand here is that of the Deepfakes technology providing grounds to deny the authenticity of a video. Examples of such plausible deniability being invoked can be found as a comment to a video showing WikiLeaks persona Julian Assange's last statements before a 7-month communications ban at the Ecuadorian Embassy in London. A Reddit user claimed that this video, and all other recent videos of Assange, were fake (Field notes July 19, 2018). The context is seems to be the widespread theory of a conspiracy to cover up Assange's alleged death. Similar sentiment was expressed in commentary to videos of Donald Trump using a derogatory term that he himself had claimed never to have used. One user commented that all the video evidence had been faked, while another linked to a Technology Demonstration Deepfake and reiterated its implications, effectively denying the Trump video's authenticity (Field notes September 10, 2018).

In an unrelated Trump discussion, a user referred to the media attention of Deepfakes and related technologies, pointing out that the narrative of fake videos had become very apparent (Field notes July 19, 2018). His guess was that a faux incriminating video of Donald Trump had been faked by Hillary Clinton, and was about to be released, while Donald Trump was preparing for the fallout by (legitimately) pushing the idea of false videos. These anecdotes show how the idea of the Deceptive Deepfake can be molded to deny the authenticity of videos, both before and after the fact.

The recurring theme through most of the public attention around the Deepfakes phenomenon and technology does not only invoke the idea of post-truth, but also implies that Deepfakes primarily is a tool to be utilized in scarily destructive ways by those so inclined. This narrative does not come as a surprise to those well versed in the history of media technologies:

> In the vulnerable phase of new media implementation, the risk of authenticity scandals, or miscommunication, is particularly high. The launch of new media technologies often facilitates new ways of both representing and manipulating reality, which in turn challenges established audience practices for interpreting the media. Authenticity scandals might result from audience reactions, media coverage, or policy reactions – in whatever case, the reactions will be significant. (Enli, 2015, p. 132)

Enli might as well have had the Deepfakes in mind when she detailed the effect of new media technologies, as it seems to describe the Deepfakes phenomenon quite accurately. Audiences are aware that videos not always represent reality, but we are only used to certain kinds of fakery. When we see a video showing a person, it is normal to assume that the face does indeed belong to the body. With Deepfakes, this is no longer necessarily the case. It is a new way of manipulating reality, which judging by the media narrative does challenge audience practices. In this case, the authenticity scandals haven't yet fully come into play, although the narrative does suppose that they will, and the history of new media does support the possibility (McLuhan 1995/1964 in Enli, 2015).

Deepfakes is not the only machine learning technology of its sort (chapter 3.1.1). Like with Deepfakes, the potential of Face2Face was demonstrated in videos making their rounds on the web, e.g. Barack Obama doing a fake speech. Commentators described the implications of Face2Face in similar fashion to what we are now seeing in regards to Deepfakes. While Face2Face gained some traction in mainstream media, there was one significant difference: it was not described as a tool available to anyone. In fact, the researchers behind the technology were frequently quoted on the fact that they chose not to make the software or source code publicly available. This is not just a key difference in the reporting on the two technologies, but also a key difference in their continuing development. In the following parts I will describe how the Deepfakes technology had some institutional roots, but still did not exist as a phenomenon until it turned public.

3.2 The technology

3.2.1 The technological system of Deepfakes

Would the Deepfakes phenomenon have existed if u/deepfakes did not share the Deepfakes code? u/deepfakes seems to argue that this contribution did not matter much in the big picture. Their opinion seems to be that machine learning by now is an inevitable development across many fields. They could lend support for this view from history that shows how technologies have often been developed simultaneously by different parties or with help from several contributors. The inventors that inventions are attributed to aren't necessarily as independent as the stories might make them seem (Hughes, 1986, p. 58). Perhaps particularly so when the technology in question is heavily based on other technological advances. This is also very much the case for the deepfakes algorithm. u/deepfakes' first contribution was 228 lines of code. While an additional 170 lines of the original Deepfakes code were licensed from other coders, it also required the following dependencies—code libraries—that had to be downloaded separately to be able to run the code:

- Python—an open source programming language and interpreter software
- OpenCV—an open source computer vision and machine learning software library
- Tensorflow—an open source machine learning framework (originally coded by Google)
- Keras—an application programming interface (API) for high-level neural networks

These are large code libraries that rest on a lot of research and code work. The Tensorflow library alone consists of almost 2 million lines of code at the time of writing. The Deepfakes code also requires a recent consumer computer with a powerful graphics processing unit (GPU), most often used for video games or the mining of crypto currency. These software libraries and hardware foundations are absolutely essential to the possibility of creating a Deepfake, even if neither of them were created specifically with Deepfakes in mind.

Even if only small parts of the dependencies are used, u/deepfakes' 228 lines seem to represent a minuscule amount of work in comparison the work that has gone into coding the dependencies. u/deepfakes also confirms this view when characterizing their work as "embarrassingly simple". The Deepfakes phenomenon could not have happened without its machine learning library dependencies, and we might speculate that it inevitably would have appeared in some form even with u/deepfakes, because their contribution mainly combines existing technologies in a new way. However, this is not to say that u/deepfakes' contribution is without consequence. It seems these 228 lines specifically, and the combination of previous technological progress that they invoke, has spurred the entire Deepfakes phenomenon and its public nature. While the last few years have seen several machine learning technologies capable of manipulating faces (or even voices) in videos (e.g. Mukhopadhyay et al., 2015; Pham et al., 2018; Thies et al., 2016; Vougioukas et al., 2018), the public nature of the Deepfakes technology makes it stand out in the crowd. The difference between software development done privately to create software contained in closed settings, and software development done in collaboration by Free Software or open source practices, has some important implications. During 2018 the Deepfakes phenomenon has become such a public phenomenon in several ways, as we will see in the following.

3.2.2 Deepfakes as a Free Software phenomenon

As we have seen, the Deepfakes code first surfaced on the community r/deepfakes, that has been officially removed, but still remains in part through Internet archives. The fragments of r/deepfakes that I found were largely incomplete, showing only a selection of the most popular headlines and occasionally entire discussion threads, from certain points in time. It does, however, clearly show that in the weeks following the initial Deepfakes code release, several other Reddit users had experimented with the code. The majority seem to have applied it to replace the faces in pornographic videos with those of Hollywood actresses or other celebrities. There were also frequent technical discussions, and even users already contributing new code to enhance the process of doing a face-swap. Furthermore, community members quickly suggested that the code be made into a collaborative project to be improved by several coders. To this, the original author u/deepfakes remarked: "I dont [sic] mind other people to host my code on GitHub. I don't want do it myself." On December 16th 2017, following u/deepfakes' acceptance, several users had individually uploaded the code to GitHub. GitHub is one of the dominating code sharing and version control services available, offering tools for a lot of the core practices of Free Software. One of those who shared the code on GitHub was the GitHub user deepfakes who created the code repository deepfakes/faceswap. While this user bore the same name as the original coder, on December 19th they remarked:

> Disclaimer : I am not 'deepfakes'. If you are deepfakes and want to get the ownership of the repo, let me know, I'll tell you how to do (github.com/deepfakes, 2017).

In other words, while the development continued to be associated with the deepfakes name, the original author was seemingly no longer directly involved. During 2018 the original code grew and divided into several GitHub forks. These forks are different repositories of code, representing branches where a copy of the original code is adapted by different contributors.

Deepfakes started turning Free Software when the Deepfakes code was uploaded to GitHub and coders started contributing. Although it wasn’t explicitly licensed as Free Software at this time, through the action of sharing and the settings of the GitHub fork, the code was implicitly a part of Free Software practices. While the software continued to grow in its Free Software form, there was also a closed source release of a compiled application under the name FakeApp, compiled by an anonymous author. Up until then, creating Deepfakes had not been possible without certain programming skills. FakeApp was the first attempt at making this a user-friendly experience by use of a graphical user interface (GUI) that allowed the user to load source material and create a video without writing any code or commands at all. This step toward simplicity seems important for the possible diffusion of the technology into society at large. It also served as an important point in a lot of the media attention that the technology was getting at the time.

FakeApp had a user interface made from scratch, but it was still based on the code from u/deepfakes and other contributors for the actual work that happened under the hood. After the initial release of FakeApp the open communities dedicated to creating Deepfakes were largely focusing on this specific software. In February of 2018, the FakeApp author(s) decided to bundle a version of the software with a component that utilized processing power on the end users’ computers to perform crypto currency mining for the author(s). This move got a poor reception from the Deepfakes community, which increasingly started shying away from closed source potentially containing malicious software packages that were hard for the community to detect.

Coincidentally or not, later in February the GNU General Public License v3.0, a formal licensing document, was added to the GitHub fork deepfakes/faceswap. This explicitly defined it as Free Software, and secured its future status as such. An open source graphical user interface (GUI) package called OpenFaceSwap was also soon released, offering the same simplicity to the end users that FakeApp had, while maintaining a transparency of the code involved. This effectively ended up rendering FakeApp obsolete. OpenFaceSwap was based on the deepfakes/faceswap repository on GitHub, albeit with its own GUI. Since then, a GUI has also been added as part of the deepfakes/faceswap repository on GitHub.

deepfakes/faceswap has since become the dominant fork of Deepfakes code. By October 2018 the official number of coders contributing to the repository had risen to 43. The majority of contributors had added or changed less than 100 lines of code, while the top three coders had added or changed tens of thousands lines of code. This repository had become the de facto hub for the ongoing and diverse Deepfakes development. It also had its own manifesto as a part of the official "README" – a document serving as both a description of and a guide for the project. Most of the manifesto was committed to the repository by a single user in June 2018, and largely described ethical guidelines for the project. Four bullet points in the middle of the document served as a summary:

- Faceswap is not for creating porn
- Faceswap is not for changing faces without consent or with the intent of hiding it's [sic] use.
- Faceswap is not for any illicit, unethical, or questionable purposes.
- Faceswap exists to experiment and discover AI techniques, for social or political commentary, for movies, and for any number of ethical and reasonable uses.

(deepfakes/faceswap, 2017/2018)

The manifesto is obviously written out of a need to distance the coding project from some of its uses, and from the public controversy sparked by the Deepfakes phenomenon. Yet, of course it cannot directly prevent such uses. Neither the developers nor the users of the software project explicitly belong to a group, and do not in any way have to subscribe to the manifesto. Nevertheless, the manifesto does serve as an attempt to unify the developers around a common idea.

There is a significant movement, particularly within computer science circles, concerned with the open sharing and improvement of code (chapter 2.1.2). This movement is even at a point where its principles are being applied way beyond the domain of computer science, and where it could be described as an ideology among its members making the related practices a second nature. The reflex of openly creating and maintaining infrastructures is sufficient explanation for anyone subscribing to the recursive public of Free Software (explicitly or implicitly) to share any type of code. Whether or not u/deepfakes had a specific reason to share the initial Deepfakes code is speculation at this point, but it might very well have been just the natural thing to do. In turn, the community's immediate response was to turn it into a Free Software project, at first implicitly through practices of sharing and collaboration, and then explicitly through the license. By then, it seems the ideology of the recursive public applied, and its members were quickly engaged in improving and defending their common infrastructure of code.

This history of development shows that the Deepfakes phenomenon fits well within Kelty's idea of Free Software. All five of his suggested core practices (chapter 2.1.2) are represented: the sharing of source code, the systemized way of forking different development paths, the formal licensing, the coordination through GitHub and the self-reflecting nature of the manifesto.

The fact that the technology is available for anyone to inspect, improve or simply play with is highlighted—by the public narrative as well as by myself—as one of its defining points. While engaging in coding is somewhat beyond the possible scope of my study, in the following I will detail my own experience with the *use* of the tools produced by the Deepfakes movement, to see in more detail how they actually work and feel.

3.2.3 Creating a Deepfake

To enrich my knowledge of the Deepfakes phenomenon as much as possible I have experimented with creating my own Deepfakes. To this end, I built a new computer from scratch with components suited for the task, specifically a GeForce GTX 1080 graphics processing unit, capable of machine learning tasks based on the Deepfakes requisites. As for the software side, I decided to specifically focus on the OpenFaceSwap package after reading about the many advantages compared to the original closed source FakeApp. While the software has a graphical user interface that is quite straight forward, it does require a bit of knowledge on how to set the software up correctly, and experimentation how to use the different parts of the software. This is a form of software tinkering that I am familiar with both career-wise and from my personal interests in computing. That is to say, my experience is not that of someone completely without technical skills, but of someone already rather well versed in computing.

The use of the software was rather extensively covered by different community made guides detailing the steps necessary, but the guides also took certain technical skills for granted. These guides told me to install several prerequisites, namely Microsoft's Visual Studio Redistributable and NVIDIA's CUDA and cuDNN frameworks. The frameworks Google Tensorflow, Keras and OpenCV, which the algorithm also relies on, came bundled with the OpenFaceSwap application itself. Once OpenFaceSwap was set up correctly, I had to choose two subjects that I could test out a swap with. From guides I had read and examples I had seen it seemed that subjects that look alike would give the best results. Furthermore it would be preferable if it was easy to find high quality pictures and videos of the subjects. After a quick brainstorming I decided to try out a swap involving NATO Secretary General Jens Stoltenberg and French president Emmanuel Macron. I deemed their facial shapes to be similar enough for a first attempt, and video material of both was easy to obtain.

I downloaded some videos of Stoltenberg and Macron from YouTube and tried feeding them to OpenFaceSwap. The very first attempt turned out to be less than satisfactory in terms of looking authentic. After trying to understand the process in more detail, and looking at guides and forum discussions, I identified some helpful steps that the app does not do automatically. I defined a new strategy, started anew, and performed the following steps:

1. Sliced out parts of the videos containing only the correct persons using OpenShot (a separate open source video editor)
2. Extracted still images from the videos using OpenFaceSwap
3. Extracted smaller images of faces from the still images using OpenFaceSwap
4. Manually deleted images that were not recognized correctly, such as faces being recognized where there were none, or faces being recognized upside-down
5. Used the extracted face images to create positionally aligned images using OpenFaceSwap
6. Adjusted colours of aligned images to help the model training’s facial recognition recognize the important differences
7. Used the aligned images to train a model with OpenFaceSwap by leaving it running for several nights

8. Used the trained model to create new face images (where facial features had been swapped out) using OpenFaceSwap
9. Merged face swapped still images together to create video using OpenFaceSwap
10. Evaluated results and replaced parts of source material as necessary, repeating all the steps

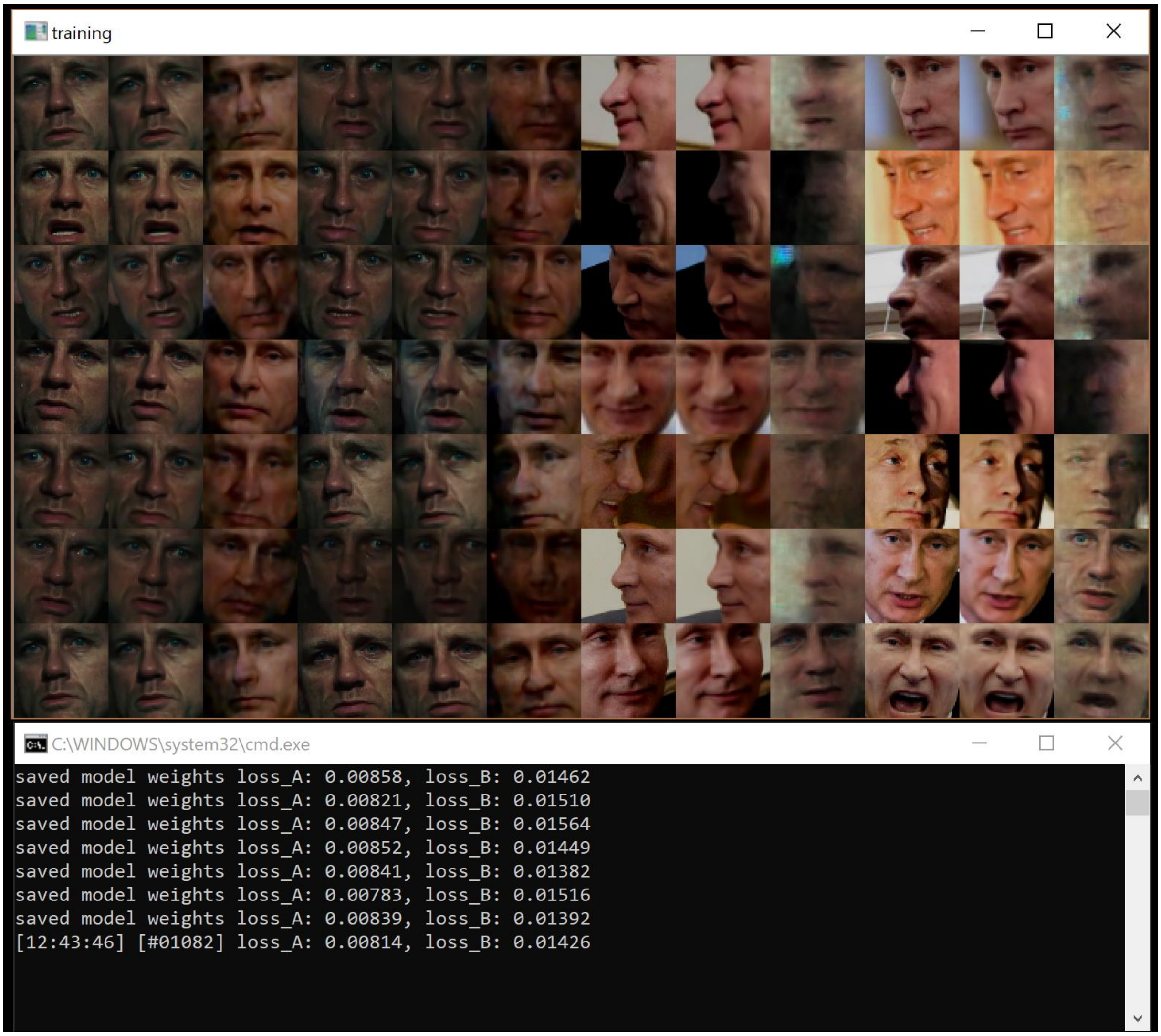
After putting many hours of work in the first set of attempts, I still ended up with a not very convincing result. I couldn't help but feel disappointed, even if the goal of my research really wasn't to end up with a perfect face-swap. From looking at the results and identifying which parts were unconvincing I could see some possible reasons that the video didn't turn out too good. It seemed simply that carefully choosing the source material is much more important than I had thought originally. Parts of the video where the source material was grainy or without detail gave less satisfactory results. Not only did the source videos and images have to be high quality and clear, but it seemed that they had to be tailored as much as possible to teaching the algorithm the correct conversions. Specifically it seemed that the imagery supplied to the algorithm had to match as much as possible in skin tones and poses. Furthermore, it would be preferable if the subjects had similar hair, forehead and ears, since the algorithm only replaced the rest of the face. After keeping this in mind when choosing new source material, editing them where necessary, and running the algorithm again for several nights in a row, the result turned out much better. The joy I felt seeing my model improving was stronger than anticipated.

Whenever I set the computer to run the model training algorithm (step 7) I was presented with two things: a preview window consisting of snapshots of conversions from face A to face B where I could follow the progress and performance, and a command line window showing the running commands and results from each iteration of the training, in numbers (Figure 6). Each morning when I woke up the very first thing I would do would be to check the current conversion metrics and the conversion snapshots to see what progress my personal robot worker had done during the night. It felt really good to see that the rather costly (but still consumer grade) computer I had invested in and built from scratch was paying its dues by improving my work for me.

When I finally decided that my project replacing Jens Stoltenberg with Emmanuel Macron was as good as it would get, I decided to try again with new source material. So far many of the videos circulating in the community had been made using Donald Trump as one of the subjects, so I thought I would instead try my hand at a swap with Vladimir Putin. Since the Deepfakes algorithms are focused on replacing visual aspects only and leaving any audio untouched, I tried to find a video where the original audio would not immediately break the deception. I found a trailer for an upcoming parody movie with Finnish actor Kari Ketonen in the role of a Vladimir Putin parody, and decided that the dance moves in the video was excellent material for a faked video using the face of Vladimir Putin.

Replacing Ketonen with Putin showed to be a burdensome task in several ways. The first objective was to extract faces from videos of Kari Ketonen. At that point, almost an entire day was spent troubleshooting why the algorithms in the OpenFaceSwap tool refused to recognize any faces in the videos at all. After unsuccessfully trawling the web for answers and trying to convert frames from the videos to different formats, I was forced to give up. It seemed this was a problem that many users stumbled upon, without any clear answer as to why. Therefore, I yet again had to acquire new source material to feed to the algorithm. Specifically I looked for more source material of the Finnish actor. After having spent days preparing the source material, the faces of Ketonen and Putin, I ran the model training algorithm five nights in a row. At that point, the conversion snapshots and conversion metrics implied that the final video would be less than satisfactory. Judging by the results, the original video of Kari Ketonen dancing was simply too dark and had too much motion, resulting in still images that were too grainy for the algorithm to be able recognize and replace the faces in a convincing way throughout the video. It seemed no amount of work would make up for the poor source material, and yet again, I was forced to consider changing my target and source material.

I decided to keep working with Putin, as the source material available for him was quite good. After making searches for lookalikes, I came across side-by-side comparisons of Putin and actor Daniel Craig showing great similarities. There is also ample high quality source material available for Craig. I could have chosen any scene from a Daniel Craig film and inserted Vladimir Putin, but I did not just want to see if it was at all possible to swap them. I also wanted to see if I could find a scene that would produce an interesting swap—where the swap could have taken on a meaning of its own. The first attempt with Putin and Craig was to replace Craig with Putin in an advertisement for the London Olympics featuring both Queen Elizabeth II and Craig together. This attempt at putting Putin together with Queen Elizabeth turned out better than the previous swaps, but still not entirely convincing—again because the extracted Craig faces were somewhat grainy. I decided to look for another Craig source, and came across the film Defiance in which Craig portrays a Belarusian partisan fighting the German occupation of Belarus during World War II.



```
saved model weights loss_A: 0.00858, loss_B: 0.01462
saved model weights loss_A: 0.00821, loss_B: 0.01510
saved model weights loss_A: 0.00847, loss_B: 0.01564
saved model weights loss_A: 0.00852, loss_B: 0.01449
saved model weights loss_A: 0.00841, loss_B: 0.01382
saved model weights loss_A: 0.00783, loss_B: 0.01516
saved model weights loss_A: 0.00839, loss_B: 0.01392
[12:43:46] [#01082] loss_A: 0.00814, loss_B: 0.01426
```

Figure 6: Status window showing a Craig-to-Putin (and Putin-to-Craig) model in early training.

The film features Craig speaking with an Eastern European accent that could fit a clip of Putin better than a purely American or British accent would. It also has clips of Craig engaged in gun violence. Even without analyzing the potential implications of such a clip in full, it is not a stretch to say that it would be an interesting position to see Putin in. Deepfakes critics are likely to have imagined this type of scene. The prospect of successfully faking it, and how this could affect the conclusion of my thesis, was exciting. I very much wanted to succeed.

However, after running through the correct steps to the best of my Deepfaking ability, I had yet again fallen in the trap of choosing source material unsuitable for an easy face-swap. Many a night of algorithm training yielded nothing more than a curious video of a rather apparent Putin impersonation. At this point, I felt that my empirical work had sufficiently informed some careful conclusions. The weeks of work that I put into tinkering with the tools did not produce any video that could be deemed entirely convincing. My understanding is that there are several reasons why this is the case. For one, I tended to choose difficult source material that I thought would make an interesting swap, not just source material that would produce good results and serve as a positive demonstration of technology’s potential. Secondly, while excited about the possible result, my goal was primarily to explore the technology, not specifically to end up with a convincing video. Thirdly, a single semester offers limited time available to work with the technology. Nevertheless, I am able to conclude that my experiments suggest the creation of a convincing Deepfake is no easy task, even if the necessary tools are easily available. That is not to say that creating a convincing fake is not possible given sufficient efforts. Some of the Technology Demonstration Deepfakes found in the wild suggest that it may be. Given this context, it is time to look more closely at what kinds of change the idea of the Deceptive Deepfake may be a part of.

3.3 The main analysis

3.3.1 The social imaginary of a visual post-truth

Deepfakes are commonly associated with the idea of a post-truth world (chapter 3.1.3). The reasons for this association is quite easily spotted when we compare the practices we have detailed so far to Sismondo's five tropes (chapter 2.1.2). Especially so when we consider the idea of the Deceptive Deepfake. The Deceptive Deepfake is by definition not based on fact (trope 1); it can be tailored to confirm the beliefs of parts of the public (trope 2); its idea can be used as plausible deniability of video proof (trope 3); it can be tailored to stir up fear and hatred without factual grounds (trope 4); and it can reduce the trust in video media and video proof (trope 5).

One implication of the characteristics of the Deceptive Deepfake is that it is capable of creating new truths. Other types of Deepfakes are also relevant to creating truth. In light of the distinction between fact and truth (chapter 2.1.1), when Putin is merged into the shape of Dr. Evil in a Satirical Deepfake it doesn't scientifically claim as a *fact* that Putin is evil or that Donald Trump is his lackey. It might however call forth the same as a *truth*, in a social and not-at-all scientific way. A social truth could also be created if a faked video plays into someone's confirmation bias (Koslowski & Maqueda, 1993; Kuhn, 1970; Wason, 1960), overpowering its lack of authenticity. There are social reasons that these truths are held to be true, and they cannot necessarily always be combated by pointing out that the videos have been tampered with. Thus, while Deepfakes might not constitute knowledge production in the sense that it produces valid knowledge in the scientific sense, it certainly does constitute a potential tool of knowledge production in the post-truth sense—where knowledge doesn't have to correspond to a singular reality. Satirical Deepfakes and Meme Deepfakes can be used to reinforce truths, and, perhaps more importantly, the idea of the Deceptive Deepfake paints Deepfakes as capable of creating truths based on faked facts (alt-facts). Since the technology is available to the public, this form of knowledge production is no longer reserved for institutional authorities.

Creating a credibly faked video, a Deceptive Deepfake, is in my experience certainly not an easy task yet—but it still seems plausible. Like any form of post-truth, it is hard work (chapter 2.11). Furthermore, whether one accepts it as plausible or not, the fact remains that the popular narrative surrounding Deepfakes paints it capable of creating truths in the hands of whoever wields it (namely anyone, including both the public and institutions of power). This is where the *idea* of Deepfakes and the *idea* of post-truth come together as mutually relevant. Whichever definition of post-truth we subscribe to; the two ideas effectively reinforce each other. Through the idea of post-truth, we are made well aware of the fact that there are powers at work creating deceptions meant to influence public opinion. It is natural to expect these powers will use any new technological possibilities, like Deepfakes, to their advantage. Looking at the other way around, Deepfakes are also channeling the idea of post-truth, through visual, tangible demonstrations of how truths can be constructed. The Technology Demonstration Deepfake (channeling the idea of Deceptive Deepfake) is almost the perfect poster child to convince someone that post-truth is real. Thus, the perceived importance of Deepfakes relies on the acceptance of the post-truth threat, and the post-truth threat seems to grow as we start to consider technologies like Deepfakes coming into play.

At this point, we should return to Taylor's social imaginary – a merging where material practices and ideas come together as an inseparable whole. The co-production of and mutual reinforcement shared by the idea of post-truth and the idea of Deepfakes can be seen as such a social imaginary. Taylor (2002) employs the term social imaginary to describe the entire commonly accepted idea of the modern society. In comparison, my use here is much more limited, and the suggested social imaginary of Deepfakes is merely an emerging one, not necessarily an accepted one. Still, I think the term is helpful in describing the way the imagined Deceptive Deepfakes are part of the idea of post-truth: The actual practice of using Deepfakes so far seems quite far removed from the frequent warnings of potentially evil applications, but the practices nevertheless support the popular Deepfakes narrative. In turn, they conjure a social imaginary in which the available infrastructure effectively allows anyone to produce truths through imagery—thus rendering video an untrustworthy media. This social imaginary seems to be a transformation of a previously rendered social imaginary of the post-truth condition where video was not yet a significant component. This transformation has added Deepfakes as an example of post-truth in practice. In Taylor's idea of the social imaginary, people take up, improvise or are inducted into new practices when a theory is formed. "These practices are made sense of by the new outlook, the one first articulated in the theory; this outlook is the context that gives sense to the practices" (Taylor, 2002, p. 111). This informs how the idea of the Deceptive Deepfake makes sense of the public's experimentation with creating one—my own included. My interest in the attempted creation of a Deceptive Deepfake relies on the fact that the idea of the Deceptive Deepfake had already been formed. It's as if the idea wills the practice into being. Similarly we can explain the ongoing work of creating tools for detecting future Deceptive Deepfakes (chapter 3.1.3) as a practice that confirms the idea of the Deceptive Deepfake. All of these highlighted practices near assume the reality of Deceptive Deepfakes. Effectively then, I argue that a change has occurred. The shifting social imaginary constitutes a reorientation towards a post-truth that also encompasses visual post-truth. The idea of post-truth has been extended to contain ideas of how videos can be forged using Deepfakes.

This reorientation is not necessarily a permanent one. In Lippmann's argument of a pseudo-experience of having seen or heard mediated facts (chapter 2.1.3), we could replace 1920s news media with Deepfakes and we would end up with the narrative in today's reporting on Deepfakes. In other words, the components of post-truth are not entirely new phenomena. The effects that are associated with the introduction of the Deepfakes technology is a recurring theme with new media technologies. Enli suggests that authenticity scandals due to new media be understood in cyclical terms. The triggers and processes of negotiation have throughout history been fairly similar from one cycle of new media technologies to the next, and it seems to be an expected result of introducing new practices (Enli, 2015, p. 136). However, after a time, the relationship between producer and audience usually stabilizes and a new authenticity contract (chapter 3.1.2) is negotiated (Enli, 2015, p. 136). Whether or not that will be the case with Deepfakes remains to be seen, but at the very least the current exposition of the phenomenon is heightening the awareness of forgery. Nevertheless, if we borrow a small part of Fuller's understanding of the post-truth condition, we can argue that this move toward a visual post-truth has changed the rules of the power game ever so slightly for a time, as the next chapter will considerate.

3.3.2 Deepfakes as a tool of the recursive public

The idea of Free Software as a recursive public, does not only inform our understanding of where Deepfakes is coming from (chapter 3.2.2). It can also directly inform the interpretation of what kinds of change Deepfakes are associated with:

> Are Habermas's pessimistic critiques of the bankruptcy of the public sphere in the twentieth century equally applicable to the structures of the twenty-first century? Or is it possible that recursive publics represent a reemergence of strong, authentic publics

> in a world shot through with cynicism and suspicion about mass media, verifiable knowledge, and enlightenment rationality? (Kelty 2008, p. 23)

Kelty suggests that understanding Free Software phenomena as recursive publics allows us to see them as movements of democratization. His idea is that the modifiability of Free Software in general assists in the reorientation of power and knowledge. The same argument can be made for the modifiability of video. The video modifications that now can be done using Deepfakes algorithms have previously been exclusive to those few who had the necessary means—who had access to very specialized knowledge, tools and resources—whether they belonged to a scientific community, a media organization or a governmental institution. Now they are available to anyone—to the recursive public. Truths can be made by the recursive public, by creating alt-facts using Deceptive Deepfakes. Truths can also be reinforced through other kinds of deepfakes, like Satirical Deepfake. The power and knowledge exercised through video has been further reoriented, just as the power and knowledge exercised through media in general have been reoriented over the last forty years (Kelty 2008, p. 6). This is not something to be taken for granted, as many media technologies are created and used without any public access, at least for the most crucial first period of time where the risk of authenticity scandals are the highest.

The Deepfakes phenomenon could thus be understood as the recursive public of the Free Software movement effectively (whether consciously or subconsciously) fighting to shift the power of knowledge production, while institutions of power, like the US Department of Defense are working to uphold the status quo. Interestingly, it is specifically the Deepfake production techniques that are developed openly by the public, while Deepfake detection techniques on the other hand are developed by institutional experts. It is as if both parties have implicitly agreed on a division of labour. When researchers develop methods of detecting the manipulation of videos (chapter 3.1.3) they are effectively creating for themselves (and possibly for their US Defense Department contractors) an expertise in video authenticity.

4 Conclusions

The Deepfakes technology is one of many machine learning technologies with similar applications. While these similar technologies are mainly researched and used within institutions of power, the Deepfakes technology was born publicly and explicitly turned into a Free Software phenomenon whose development is currently ongoing. The technology is being used—primarily by the public—to forge pornography, create memes, do satire and perform technology demonstrations. The imagined Deceptive Deepfake has not yet materialized, and at this point there seems to be a significant gap between being able to *imagine* a faked video and actually being able to convincingly fake it. However, the creation of a Deceptive Deepfake is not implausible, given the right conditions and resources. Nevertheless, the *idea of* the Deceptive Deepfake is already associated with new social practices no matter its actual existence. The notion of a *social imaginary* has been particularly helpful in understanding this association. A social imaginary consists of world views and social practices that reinforce each other. In the social imaginary examined herein, anyone can create or modify truths by producing Deceptive Deepfakes. This understanding of Deepfakes is essential to the publicly expressed connections between Deepfakes and the idea of *post-truth*. Due to this association, central tenets of the post-truth condition are being reinforced by the idea of the Deceptive Deepfake. Conversely, the fears and interests surrounding the Deepfakes phenomenon is being reinforced by the notion of post-truth. This is the core of the social imaginary of Deepfakes.

Itself a source of great interest and heated debate within the field of STS, the post-truth condition has paved the way for STS to see any practice capable of calling forth truths as tools of knowledge production. The Deepfakes phenomenon could thus be understood as such a tool, which the *recursive public* of the Free Software movement is using to effectively shift the power of knowledge production. This shift is a democratizing one, yet not automatically a welcome one. Conversely, institutions of power are working to uphold the status quo. This is the ongoing battle of the *visual post-truth*, that seems to have been induced not simply by the technology itself, but by the more complex social imaginary it is a part of.

Interestingly, the idea of Deepfakes seems to have a faint tinge of prophecy that could end up either as self-fulfilling or self-defeating. The attention to the idea of Deceptive Deepfakes seems to have sparked actions that could end up confirming it (like the rapid development of the technology) as well as actions that could defeat it (like the development of detection tools and the public awareness of forgery). While I have found that Deceptive Deepfakes are currently no more than an imagination or a possibility, my message is not intended at dismissing the fears associated with Deepfakes. The dominating public narrative of dangers to democracy is certainly one to take seriously. This has also been a motivation for the thesis. Following my analysis I would also like to expand the narrative of dangers somewhat: If the recursive public consciously wishes to stay relevant when the Deepfake detectors start assigning truths, perhaps it should not only focus on production, but also on detection, lest they be left to rely on institutional expertise again. In the spirit of Dewey, I argue that this expertise is no guarantee for altruism. I would not necessarily trust the US Department of Defense to assess the authenticity of videos for me. Some public control over these kinds of tools might very well be a common good. However, I will claim that an unchecked public creation of alt-truths by Deepfakes would be no good outcome. To avoid this, it seems public awareness of deception techniques could become necessary. So could the creation of democratic detection techniques. This all depends on the further development of the phenomenon.

This has only been a first-look at a brand new and evolving technology, which should be extended and revised by other studies as the phenomenon develops. Rather than an overconfident analysis or a brazen prediction of the future, I hope to have given an insight into the complexities of Deepfakes, put the fears into context, and detailed my impression of the current socio-technical practices in a sober manner.

References

Beaulieu, A. (2010). Research Note: From co-location to co-presence: Shifts in the use of ethnography for the study of knowledge. *Social Studies of Science*, *40*(3), 453–470. https://doi.org/10.1177/0306312709359219

Burrell, J. (2009). The Field Site as a Network: A Strategy for Locating Ethnographic Research. *Field Methods*, *21*(2), 181–199. https://doi.org/10.1177/1525822X08329699

Cadwalladr, C. (2018, March 18). ‘I made Steve Bannon’s psychological warfare tool’: meet the data war whistleblower. *The Guardian*. Retrieved from https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump

Chesney, R., & Citron, D. (2018, February 21). Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy? Retrieved 12 October 2018, from https://www.lawfareblog.com/deep-fakes-looming-crisis-national-security-democracy-and-privacy

Christian, J. (2018, January 2). Experts fear face swapping tech could start an international showdown. Retrieved 17 April 2018, from https://theoutline.com/post/3179/deepfake-videos-are-freaking-experts-out

Cole, S. (2017, December 11). AI-Assisted Fake Porn Is Here and We’re All Fucked - Motherboard. Retrieved 9 October 2018, from https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn

Collins, H., Evans, R., & Weinel, M. (2017). STS as science or politics? *Social Studies of Science*, *47*(4), 580–586. https://doi.org/10.1177/0306312717710131

deepfakes/faceswap. (2018, October 11). deepfakes/faceswap/README.md. Retrieved from https://github.com/deepfakes/faceswap/blob/master/README.md (Original work published 19 December 2017)

Dewey, J. (1929). The Public and its Problems. *Journal of Philosophy*, *26*(12), 329–335.

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing Ethnographic Fieldnotes, Second Edition*. The University of Chicago Press. Retrieved from http://www.press.uchicago.edu/ucp/books/book/chicago/W/bo12182616.html

Enli, G. (2015). *Mediated authenticity: how the media constructs reality*. New York: Peter Lang.

Fikse, T. D. (2018). *Understanding digital STS: matter and methods* (Exam). University of Oslo.

Fuller, S. (2016, December 25). Embrace the Inner Fox: Post-Truth as the STS Symmetry Principle Universalized, Steve Fuller. Retrieved 3 April 2018, from https://social-epistemology.com/2016/12/25/embrace-the-inner-fox-post-truth-as-the-sts-symmetry-principle-universalized-steve-fuller/

Fuller, S. (2017, April 1). Is STS all Talk and no Walk? Retrieved 14 October 2018, from https://easst.net/article/is-sts-all-talk-and-no-walk/

Fuller, S. (2018). *Post-Truth: Knowledge As A Power Game*. Anthem Press.

Geertz, C. (2000). *Interpretation of Cultures*. New York, NY: Basic Books. Retrieved from http://www.aspresolver.com/aspresolver.asp?ANTO;1667767

github.com/deepfakes. (2017, December 19). Dead project ? Let’s move on · Issue #12 · joshua-wu/deepfakes_faceswap. Retrieved 7 July 2018, from https://github.com/joshua-wu/deepfakes_faceswap/issues/12

Hammersley, M., & Atkinson, P. (1986). Ethnography: Principles In Practice. *Canadian Journal of Sociology / Cahiers Canadiens de Sociologie*, *10*. https://doi.org/10.2307/2070079

Hawkins, D. (2018, February 8). Reddit bans ‘deepfakes,’ pornography using the faces of celebrities such as Taylor Swift and Gal Gadot. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/morning-mix/wp/2018/02/08/reddit-bans-deepfakes-pornography-using-the-faces-of-celebrities-like-taylor-swift-and-gal-gadot/

Hine, C. (2015). *Ethnography for the Internet: Embedded, Embodied and Everyday*. Bloomsbury Publishing.

Hsu, J. (2018, June 22). Experts Bet on First Deepfakes Political Scandal. Retrieved 9 October 2018, from https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/experts-bet-on-first-deepfakes-political-scandal

Hughes, T. P. (1986). The Evolution of large technological systems. Wissenschaftszentrum.

Jasanoff, S., & Kim, S.-H. (Eds.). (2015). *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power* (1 edition). Chicago ; London: University of Chicago Press.

Jasanoff, S., & Simmet, H. R. (2017). No funeral bells: Public reason in a ‘post-truth’ age. *Social Studies of Science*, *47*(5), 751–770. https://doi.org/10.1177/0306312717731936

Johnson, D. B. (2018, July 16). Rubio warns on ‘deep fakes’ in disinformation campaigns -. Retrieved 9 October 2018, from https://fcw.com/articles/2018/07/16/deep-fakes-rubio-warner.aspx

Keane, S. (2018, September 5). Congress wrestles with ‘deepfake’ threat to Facebook. Retrieved 9 October 2018, from https://www.cnet.com/news/congress-wrestles-with-deepfake-threat-to-facebook/

Kelty, C. M. (2008). *Two Bits: The Cultural Significance of Free Software*. https://doi.org/10.1215/9780822389002

Knight, W. (2018, August 7). The Defense Department has produced the first tools for catching deepfakes. Retrieved 9 October 2018, from https://www.technologyreview.com/s/611726/the-defense-department-has-produced-the-first-tools-for-catching-deepfakes/

Koslowski, B., & Maqueda, M. (1993). What Is Confirmation Bias and When Do People Actually Have It? *Merrill-Palmer Quarterly*, *39*(1), 104–130.

Kuhn, T. S. (1962). Historical Structure of Scientific Discovery. *Science*, *136*(3518), 760–764.

Kuhn, T. S. (1970). *The structure of scientific revolution.* Chicago: University of Chicago Press.

Li, Y., Chang, M.-C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *ArXiv:1806.02877 [Cs]*. Retrieved from http://arxiv.org/abs/1806.02877

Lippmann, W. (1922). *Public opinion*. Macmillan,. Retrieved from http://hdl.handle.net/2027/uc1.b5232744

Lippmann, W. (1925). *The Phantom Public*. Transaction Publishers.

Lynch, M. (2017a). STS, symmetry and post-truth. *Social Studies of Science*, *47*(4), 593–599.

Lynch, M. (2017b, February 6). Post-truth, Alt-facts, and Asymmetric Controversies (Part I) – First 100 Days. Retrieved 3 April 2018, from http://first100days.stsprogram.org/2017/02/06/post-truth-alt-facts-and-asymmetric-controversies-part-i/

Mukhopadhyay, D., Shirvanian, M., & Saxena, N. (2015). All Your Voices are Belong to Us: Stealing Voices to Fool Humans and Machines. In G. Pernul, P. Y A Ryan, & E. Weippl (Eds.), *Computer Security -- ESORICS 2015* (Vol. 9327, pp. 599–621). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24177-7_30

OED Online. (n.d.). post-truth, adj. *OED Online*. Oxford University Press. Retrieved from http://www.oed.com/view/Entry/58609044

Pham, H. X., Wang, Y., & Pavlovic, V. (2018). Generative Adversarial Talking Head: Bringing Portraits to Life with a Weakly Supervised Neural Network. *ArXiv:1803.07716 [Cs]*. Retrieved from http://arxiv.org/abs/1803.07716

Shifman, L. (2014). *Memes in Digital Culture*. MIT Press. Retrieved from https://www.jstor.org/stable/j.ctt14bs14s

Sismondo, S. (2017a). Post-truth? *Social Studies of Science*, *47*(1), 3–6. https://doi.org/10.1177/0306312717692076

Sismondo, S. (2017b, May 1). Not a Very Slippery Slope: A Reply to Fuller. Retrieved 14 October 2018, from https://easst.net/article/not-a-very-slippery-slope-a-reply-to-fuller/

Sørensen, E. (2017, April 1). The social order of facts vs. truths | EASST. Retrieved 4 April 2018, from https://easst.net/article/the-social-order-of-facts-vs-truths/

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, *43*(3), 377–391. https://doi.org/10.1177/00027649921955326

Taylor, C. (2002). Modern Social Imaginaries. *Public Culture*, *14*(1), 91–124. https://doi.org/10.1215/08992363-14-1-91

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on* (pp. 2387–2395). IEEE.

Vougioukas, K., Petridis, S., & Pantic, M. (2018). VOUGIOUKAS ET AL.: SPEECH-DRIVEN FACIAL ANIMATION WITH TEMPORAL GANS 1 End-to-End Speech-Driven Facial Animation with Temporal GANs.

Walker, S. (2015, April 2). The Russian troll factory at the heart of the meddling allegations. *The Guardian*. Retrieved from https://www.theguardian.com/world/2015/apr/02/putin-kremlin-inside-russian-troll-house

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. https://doi.org/10.1080/17470216008416717

Wyatt, S. (2007). Technological Determinism Is Dead; Long Live Technological Determinism. In E. J. Hackett, O. Amsterdamska, Lynch, Michael, & J. Wajcman (Eds.), *The Handbook of Science and Technology Studies, Third Edition*. Cambridge, Massachusetts: The MIT Press. Retrieved from http://www.virtualknowledgestudio.nl/documents/handbook-chaptersally.pdf