Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Social and Educational Sciences
Department of Sociology and Political Science

Regine Dotset Ringerud

Deepfakes - the future of disinformation

A quantitative study of Norwegians' ability to detect deepfakes

Master's thesis in Media, communication and information technology
Supervisor: Lisa Reutter
Co-supervisor: Pieter de Wilde
June 2021

NTNU
Norwegian University of Science and Technology

Regine Dotset Ringerud

Deepfakes - the future of disinformation

A quantitative study of Norwegians' ability to detect deepfakes

Master's thesis in Media, communication and information technology
Supervisor: Lisa Reutter
Co-supervisor: Pieter de Wilde
June 2021

Norwegian University of Science and Technology
Faculty of Social and Educational Sciences
Department of Sociology and Political Science

NTNU
Norwegian University of
Science and Technology

Abstract

Deepfakes are manipulated videos, audio, or images where machine learning is used to make them as realistic as possible. They are a relatively new phenomenon, and research on its detection is scarce, specifically in a Norwegian context. This thesis explores Norwegian's ability to detect deepfakes. The study's research question is: To what extent are Norwegians able to recognize deepfakes, and which factors affect this? The factors examined include gender, age, education, digital literacy, internet use, trust in news, interest in politics, expected performance and previous knowledge. The study uses a quantitative design where an online survey was created and distributed among Norwegian citizens, primarily using Facebook. In the survey, respondents were asked to evaluate the authenticity of 16 videos, where 8 were real and 8 were deepfakes. The survey had 682 respondents.

Theoretical terms like knowledge gaps, digital literacy, and confirmation bias, compose the theoretical framework for examining how humans treat media content, and the potential consequences of not being able to detect fake content. The central findings of the research were that Norwegian's had an average success rate of 57.3% when classifying videos, which is only slightly higher than randomly guessing. Older participants performed worse than younger participants, and hours spent on the internet had a curvilinear effect on correctly classifying videos. Expecting to perform well had a positive influence on performance and confidence and having previous knowledge of the subject in the video or the video in itself increased respondents' confidence when answering.

These results are an important addition to the research field, and they show that general awareness and detection of deepfakes in Norway is relatively low. Hence, they are a potentially powerful threat if used maliciously.

Sammendrag

Deepfakes er manipulerte videoer, lyd eller bilder hvor maskinlæring blir brukt for å gjøre dem så realistiske som mulige. De er et relativt nytt fenomen, og forskning på temaet er begrenset, spesielt i en norsk kontekst. Denne oppgaven utforsker nordmenns evne til å gjenkjenne deepfakes. Problemstillingen for oppgaven er: I hvilken grad er nordmenn i stand til å gjenkjenne deepfakes, og hvilke faktorer påvirker dette? Faktorene som ble utforsket er kjønn, alder, utdanning, digital kompetanse, internettbruk, tillitt til nyheter, politisk interesse, forventet prestasjon og tidligere kjennskap. Studien bruker et kvantitativt design hvor en online spørreundersøkelse ble laget og distribuert bland norske innbyggere, hovedsakelig ved bruk av Facebook. I spørreundersøkelsen ble respondenter spurt om å evaluere autentisiteten til 16 videoer, hvorav 8 var ekte og 8 var deepfakes. Spørreundersøkelsen hadde 682 respondenter.

Teoretiske begreper som kunnskapskløfter, digital kompetanse og bekreftelsestendenser legger et godt teoretisk rammeverk for å undersøke hvordan mennesker behandler medieinnhold, og de potensielle konsekvensene av å ikke være i stand til å gjenkjenne falskt innhold. Sentrale funn i studien er at nordmenn hadde en gjennomsnittlig suksessrate på 57.3% når det kom til å evaluere videoer, som er bare litt høyere enn ved tilfeldig gjetning. Eldre deltakere presterte dårligere enn yngre deltakere, og antall timer som brukes på internett hadde en kurvelineær sammenheng med evnen til å gjenkjenne deepfakes. Forventninger om å prestere godt hadde en positiv innvirkning på prestasjon og selvtillit, og det å ha tidligere kunnskap til personen eller videoen økte deltakernes selvtillit når de evaluerte.

Disse resultatene er et viktig tilskudd til forskningsfeltet, og de viser at på generell basis er bevissthet og gjenkjenningsevne rundt deepfakes i Norge relativt lav. Derfor kan deepfakes være en mektig trussel hvis det brukes ondsinnet.

Preface

There are quite a few people who have contributed greatly to the completion of this thesis. Firstly, I would like to give the biggest thank you to my co-supervisor Lisa Reutter, who not only gave invaluable academic advice, but was also one of my biggest supporters through constant reassurance and enthusiasm. Further, I would also like to thank my supervisor Pieter de Wilde, who helped me see the research through new eyes when my head was stuck. Moreover, I would like to thank Ane Møller Gabrielsen at NTNU University library, who helped me so much with the development of the survey. I very much appreciate all the time you set aside to solve my problems. Thank you to all participants of the survey who contributed to valuable insight into the human detection of deepfakes.

Last, but far from least, I would like to give a huge thank you to my friends and family who have always been my personal cheerleaders. Your support and love have been crucial during my time as a student in Trondheim.

Content

Models

Figures

Tables

1 Introduction

As the tools for manipulating multimedia become increasingly sophisticated and accessible, the importance of recognizing the manipulated content also increase. Multimedia manipulation has been around for a long time, and to some extent, it is now general knowledge that a photograph could be tangled with without it being visually recognizable. Simultaneously, manipulation tools have become better, cheaper, and even more comprehensive. This has led to the kind of multimedia manipulation called deepfakes. A deepfake is a hyper-realistic video that is digitally manipulated to depict people who say or do things they never did in real life (Westerlund, 2019). This is a form of manipulation that uses machine learning to exchange someone's face with another person's face, forge someone's voice and get them to say or do something they never said or did or produce a completely false audio of someone talking (Diakopoulos and Johnson, 2020; Vaccari and Chadwick, 2020). Deepfakes are a phenomenon that is thought to infiltrate our lives to a greater extent in the future, which has led the Norwegian Media Authority to develop a learning resource for high school students specifically tackling deepfakes and critical understanding of media (Krogsrud and Velsand, 2020). The resource includes assignments, questions for discussion, cases, quizzes, and a poster with tips and rules to remember when encountering a deepfake.

Furthermore, we see that deepfakes also have been more prominent in the media over the past few years and that their occurrence in normal people's everyday lives is increasing. A deepfake of Barack Obama was in 2018 created by Buzzfeed to spread awareness of how sophisticated the technology was, including a reveal at the end of the video showing that it was Jordan Peele who was the voice behind the deepfake (BuzzFeedVideo, 2018). Likewise, in March of 2021, a cheerleader-mom was accused of making deepfakes of three girls, trying to make it look like they were naked, drinking alcohol, and smoking in an attempt to frame the girls and have them kicked off of her daughters' cheerleading team to help her daughter get ahead (Elliott, 2021). Senior advisor from the Norwegian guidance service for people who have been violated online argues that this might be more common in the future and says that they receive inquiries about similar incidents regularly (Alnes, 2021).

Norway is a country with a small population and a strong social-democratic governance, where trust in government and democracy, in general, is high, and where digital literacy has been a crucial part of school-curriculum for the past 25 years (Newman et al., 2018; Erstad 2006, p. 416). Perhaps because of this, Norway's problem with distributing fake and manipulated information has been relatively low compared to several other countries like the United States, Great Britain, and Austria (Kalsnes 2019, p. 51). In a Reuters Digital News Report study, Norwegians stand out in the sample in several ways. For instance for being less exposed and less concerned by different forms of dis- and misinformation (Newman et al. 2018, p. 37). Participants from Norway also reported that they were more concerned about shoddy journalism than news articles wholly made up. Compared to respondents from the United States and European countries, Norwegians report that they come across news content that they believe is fake at a much lower rate.

Simultaneously, the Digital News Report also shows that Norwegians have a lower rate of certainty when it comes to believing in their abilities to recognize fake information. These decreased concerns for being exposed to misinformation and their low level of trust in their abilities to recognize fake content, might make Norwegians vulnerable in the face of misinformation in general, and perhaps deepfakes especially. This gives reason to expect that a Norwegian sample might show divergent results from previous research conducted by Schetinger et al. (2017), Rössler et al. (2018), or Thaler (2019). It is important to understand the extent to which Norwegians can identify fake content because not doing so might cause an increase of misled citizens, which in turn could lead to increased distrust in news outlets, the governance, and perhaps democracy as a whole (Citron and Chesney, 2019; Day, 2019). This, in turn, might lead to increased polarization in the public, where political gaps between citizens grow and might lead to hateful and violent struggles (Citron and Chesney, 2019).

Today, deepfakes are often used when making pornography, by putting someone's face (usually a female celebrity) onto a woman's body in an existing pornographic video to make the impression that the video portrays a real pornographic video of this celebrity. This technique is also often used to make revenge-porn by putting the face of an ex-girlfriend onto the body of a porn actress (Diakopoulos and Johnson, 2020; Meskys et al., 2020). Experts also fear that deepfakes might be used in political campaigns in the future as a means for blackmail and exploitation, and harassment (Franks and Waldman, 2019). Deepfakes, how they are made, and how to detect them are being thoroughly researched in technological contexts, where blockchain technology and algorithms are essential components (See for instance Fernando et al. (2019) and Hasan and Salah (2019)). However, deepfakes are an emerging object of inquiry within social sciences too, because of how the spread of misinformation affect people's critical thinking and overall trust (Hasan and Salah, 2019; Schiff et al., 2020).

Although some scientists argue that deepfakes may have numerous positive areas of application, for instance, within education, healthcare, technology, film, art- and culture, shopping, writing, and tourism, we can already see that deepfakes also have troubling properties (Meskys et al., 2020; Donovan and Paris, 2019; Westerlund, 2019; Diakopoulos and Johnson, 2020; Kwok and Koh, 2020; Silbey and Hartzog, 2019). Potential harmful applications of deepfakes that are thought to become more prevalent in the future are political attacks to mislead the public voters, blackmailing, identity theft, and cyber-terrorism (Vaccari and Chadwick, 2020; Donovan and Paris, 2019; Westerlund, 2019; Antinori, 2019). All the above will most likely contribute to a growing distrust in news outlets, politicians, and democracy in general (Westerlund, 2019). This might lead to an informational gap between the people who can identify a fake video and those who cannot. Seen in a bigger context and over time, this may create populations with a growing distrust in news media and parts of the population who cannot participate in democracy because they do not have the information needed (Day, 2019).

Deepfakes might be used in political campaigns to undermine the competitor and attract voters to their political party, especially if the politician portrayed in the deepfake cannot convince their voters that the media is fake (Vaccari and Chadwick, 2020; Schiff et al., 2020). These kinds of behaviors could have significant implications for citizens' competence, as well as the quality of a country's democracy by tampering with elections as well as compromising national security (Vaccari and Chadwick, 2020; Citron and Chesney, 2019; Mirsky and Lee, 2021).

Another harmful area of use that is thought to have implications in the future is to use deepfakes as blackmail (Westerlund 2019, p. 39). Already, we have seen one such example where Rana Ayyub, an Indian journalist, was sent an email containing a pornographic deepfake of herself in an attempt to silence her criticism of the Indian government (Ayyub, 2018). Further on, deepfakes might be used to create fraudulent identities or even identity theft.

Further on, deepfakes might lead to issues with cybersecurity, for instance, by manipulating stock markets by showing manipulated footage of a CEO saying misogynistic or racist slurs, making inaccurate statements of a company's financial loss, or announcing a fabricated merger (Westerlund, 2019). An example of this emerged in 2019 when cybercriminals used deepfake technology to impersonate a chief executive's voice demanding a transfer of approximately $243000 in a conversation with his employee (Stupp, 2019). Therefore, researching individuals' ability to identify such fake content is essential to understand better what measures should be taken to prevent the harmful spread of deepfakes that might lead to a growing distrust among populations.

1.1 Research question

Deepfakes is a relatively new phenomenon. The term was first used in 2016, so research on the topic is relatively scarce in the Norwegian context. However, misinformation and manipulation of media are not new and have been a topic of media studies for decades (Wardle and Derakhshan, 2017, Uberti, 2016). Phenomena like misinformation and media manipulation are important within media studies because of their potential to affect people's critical thinking, trust in news outlets, their overall news literacy, and the way it might affect how people treat media content (McDougall, 2019; Marchi, 2012; Scheufele and Krause, 2019). We separate misinformation and disinformation by the sender's intent, where misinformation describes fake media with no intention to deceive. In contrast, disinformation is made and spread with the intention to deceive (Wardle and Derakhshan, 2017).

The technology behind deepfakes might significantly impact society. Based on the potential areas of use, the overall challenge will be for individuals to detect deepfakes with certainty in the news stream on social media platforms. Because the technology has become so sophisticated, cheap, and easily accessible, one no longer needs to own advanced equipment or have much experience to make a deepfake. Today, they can easily be made using free apps like Reface or iface or by downloading open-source code from public websites like Github (Mirsky and Lee 2021, p. 17).

Like most research on emerging information and communication technology, research on deepfakes has been framed and explored mainly as binaries, between continuity and discontinuity and between utopia and dystopia (Boczkowski and Lievrouw, 2008). Researchers disagree on how dangerous this developing technology is and how much it differs from previous forms of multimedia manipulation. Some scientists argue that manipulation has been around for centuries and that deepfakes are no different than the discovery of photoshopping (Donovan and Paris, 2019). Other scientists consider

deepfakes to be unique in their form and hence form a different and more critical threat than other forms of manipulation (Bates, 2018).

Based on the above mentioned previous deepfake detection within the Norwegian context, I propose the following research question in this study:

> *To what extent are Norwegians able to recognize deepfakes, and which factors affect this?*

In addition to researching whether Norwegian's can detect deepfakes, I will also look at their confidence and certainty when detecting, and whether most participants are able to give an answer based on critical thinking and audiovisual cues, or if they find it too difficult and primarily give uncertain answers. Further, I will explore which factors influence this confidence.

Quantitative research was thought to be the best fitting research design, as this is a recognized method to measure the ability of deepfake detection within a population in the field (Khodabakhsh et al., 2019; Korshunov and Marcel, 2020). I will answer the research question by using an online survey developed and distributed for the purpose of this research project where participants were exposed to 16 videos, 8 of which were real and 8 were deepfakes. The videos portrayed politicians at press conferences or in interviews, actors and celebrities in commercials, interviews, or scenes in movies, and a singer playing saxophone. Participants would then determine whether they believed the video was a deepfake or not. The respondents also answered additional questions about whether they knew the subject in the video, or the specific video before or other versions of it, and what cues in the video were decisive for their decision. Demographic information like gender, age, political interest, and other questions was also included in the survey to differentiate between the target groups. The survey will determine to what extent Norwegians can recognize deepfakes and which factors might contribute to these abilities, as well as insight into their confidence when responding.

The thesis will be divided into six chapters. After the introduction, I will discuss what deepfakes are, how they are made, and how they can be used. Then I will describe the terms informational gaps, digital literacy, and confirmation bias and explain why these terms are important in the context of deepfakes. Further, I will give an in-depth description of what previous research on the topic has found before introducing the hypotheses. In the next chapter, the research design and method will be described, and a detailed description of the online survey will be provided. Then, the analysis chapter will be divided into two sub-chapters. I will first present the descriptive data and then test the hypotheses mentioned previously using both a linear and logistic regression—lastly, a discussion of the results in the context of the hypotheses and summarizing conclusion.

2 Deepfakes – a comprehensive introduction

Before going deeper into previous research and theoretical concepts in the field, it is crucial to explain the phenomenon of deepfakes and its history thoroughly. Understanding how deepfakes are made is vital to understanding why they are different from other manipulated media and why they threaten critical thinking and general trust in news outlets. Hence, this section will give a detailed introduction to the phenomenon of deepfakes and why they are important to research within social science.

Deepfakes can be described as products of AI or machine-learning operations that combine, replace, superimpose, and merge photos, video, and audio, creating a fake product of multimedia that appears to be authentic (Alexandrou and Maras, 2019). The level of sophistication in such videos may vary to a great extent. Donovan and Paris (2019) distinguish between deepfakes on the one hand, which uses complicated, sophisticated, and expensive software, and cheap fakes on the other hand, which uses software that is cheap and accessible. Cheap fakes include videos or audio that has been edited by simply using a lookalike, slowing down, speeding up, cutting, re-contextualizing, or re-staging the outtake.

According to Meskys et al., in 2020, the first recollection of what is now known as deepfakes was in a paper written by Justus Thies et al. presented at the Conference on Computer Vision and Pattern Recognition in 2016. The concept was later called deepfake after a Reddit user in 2017 used the term as a username while posting videos where female celebrities like Gal Gadot were swapped with faces of pornographic actresses (Tolosana et al. 2020, p. 132; Citron and Chesney 2019, p. 1772).

From 2017 and onward, deepfake is the term that has been used to describe these kinds of manipulated multimedia, and the technology behind the production of such deepfakes continues to develop at a rapid pace, making the detection of such deepfakes increasingly harder. Today, there are several ways that you can make deepfakes, each method having its advantages and disadvantages. However, most are based on the use of Generative Neural Networks (GAN).

Ian Goodfellow, a Google researcher, invented GAN, and the method is based on having two neural networks working against each other (Citron and Chesney 2019, p. 1760; Agarwal et al. 2019, p. 40-41). One network called a generator uses a dataset of photographs, audio, or video of the source target and produces a sample draft. The other network, called the discriminator, then judges whether the sample is of such a quality that it is convincing while learning to recognize the properties of the real video and the fake. This sequence is then repeated iteratively at a very high speed.

Even though this GAN technology is quite complicated for someone without experience with computer science and machine learning, most software is free for download on public websites like Github.com for anyone to access (Vaccari and Chadwick 2020, p. 2; Mirsky and Lee 2021, p. 24). The fact that there are also thousands or possibly millions of pictures and videos of different celebrities freely available on places like Google image search allows for everyone to experiment and create deepfakes in the comfort of their own home.

We can also see the development of this kind of technology today in our everyday lives. For instance, Snapchat, TikTok, and Instagram, who provide a Face Swap feature, allowing users to swap faces with other people, change their facial features, making them look older or younger, or make other changes to their face. Although these algorithms are not amongst the most sophisticated, they are professional enough to "learn" the features of a face from different angles, allowing them to give a precise manipulation of user's faces in all angles in real-time (Öhmam 2020, p. 133).

In recent years, reports have shown that this technology has primarily been used to create pornographic deepfakes and that a considerable amount of deepfake content online is hardcore pornographic videos. Deepfakes allow the production of sexual entertainment against the will of the video's target, and it is shown that females and queer people are disproportionately targeted by these kinds of videos (Persons 2020, p. 1; Franks and Waldman 2019, p. 894).

3 Theoretical framework and previous research

Social science research of misinformation and its detection on the internet has gained significant prominence in recent years (Freiling et al., 2021; Seo et al., 2020). This chapter will introduce and explain the theoretical concepts of knowledge gaps, digital literacy, and confirmation bias and discuss the relevance for this study. I will also introduce previous research on misinformation and deepfakes here. Knowledge gaps might contribute to a better understanding of how citizens might be affected differently by deepfakes. The term digital literacy is included to give a better understanding of what being digitally literate means and what being digitally illiterate might lead to. Moreover, the term will contribute to a greater understanding of how people consume media content and how this affects people's memory, exposure, and perception. Together this will provide the study with a sound theoretical framework and foundation for the development of the survey. I draw here from concepts and previous research in a variety of research traditions such as media science, pedagogy, political science, and psychology.

3.1 Knowledge gaps

Within media and communication theory, the theory of knowledge gaps explains the gap between those who have the knowledge and abilities necessary to utilize their available resources and those who do not (Aalberg and Elvestad, 2012). The hypothesis was first introduced by Tichenor, Donohue, and Olien in 1975 and states that mass media can increase systemic gaps in knowledge in a population, based on citizens' socio-economic status. They argued that people with higher socioeconomic status tend to have more information and acquire information faster than citizens with lower socioeconomic status (Donohue et al., 1975). Therefore, this concept is relevant in any study of new media phenomena and has previously been applied to explain gaps in the ability to recognize fake news (Gerosa et al., 2021).

Several circumstances can create such gaps, and it was previously more common to talk about knowledge gaps created by access to critical information and the lack thereof. Important information is the sort of information that allows groups of people to ensure their rights within a society, for example, their right to influence decisions through democracy. In later times, however, and especially with the commercialization of the internet, we see that today's knowledge gaps are not so much about the access to important information, but about the gap between those who can navigate the constant stream of information and select what is important, and those who cannot (Schwebs and Østbye, 2017). This gap in a population's knowledge may cause problems in health disparities and politics which might lead to differences in participatory behaviors. Not being able to navigate the enormous stream of information might cause people to not engage in preventive health resources such as cancer screenings, and a gap in political knowledge may cause gaps in political engagement, and interest (Hwang and Jeong, 2009).

According to Sande (1989), information gaps are based on the distinction between those who "knows" on a general basis and those who do not know, and this distinction is

systematically bound to social status, especially education. Additionally, those who generally have high knowledge also acquire new knowledge more easily (Sande 1989, p. 6). Based on Sande's theories, we might expect results from the survey showing that respondents with little general knowledge might perform poorly when identifying deepfakes.

Knowledge gaps are often related to education and a higher socioeconomic status because those who are highly educated often choose media outlets containing more information. They tend to choose websites, articles, and shows that are serious and have helpful content. This hypothesis about the knowledge gap might provide a better understanding of whether participants of higher socioeconomic status have an advantage when distinguishing deepfakes from genuine content. One factor in distinguishing people of higher socioeconomic status is whether they have higher education (Donohue et al., 1975). Therefore, participants in the survey were asked about their highest completed level of education to see if this affected their ability to detect deepfakes in a significant way. Based on this, we might expect participants with higher levels of education to perform well in the survey.

Another factor that may affect knowledge gaps is distrust. People tend to have great trust in friends and family, even more so than information coming from commercial actors, such as different news outlets (Aalen, 2016). Hence, if someone has distrust in news media outlets, and have greater trust in the information they get from friends and family, they might miss out on or disregard critical information that might affect their ability be productive parts of society. Hence, it is important for citizens to have some trust in news outlets, to make sure that knowledge gaps do not increase.

Those who "know" online can be described as digitally literate and those who are not can be described as being digitally illiterate (Pietrass, 2007). Further on, I will look at the term digital literacy to explain how knowledge gaps on the internet are explained through the possession of digital knowledge and lack thereof.

3.2 Digital literacy

The Norwegian ministry of education defines digital literacy as the ability to use digital tools and media in a safe, critical, and creative way (Kunnskapsdepartementet, 2017). Digital literacy revolves around performing practical tasks, communicating, collecting, or processing information. These and digital judgment in the form of privacy, source criticism, and information security are important factors of digital literacy. The term is often used within media science to research how digital literacy, or the lack thereof affects how people perceive and evaluate content online, and what the consequences of digital divides might be (Lupač, 2018; Tsai et al., 2017; Brandtzæg et al., 2011).

EU's project, DigComp, has compiled a framework of digital literacy and defines it as the ability to be aware, critical, and creative when using ICT to achieve a goal related to employment, work, learning, leisure time, inclusion, and participation in society (DigComp 2.0, 2019). Digital literacy contributes to a citizen's ability to achieve valued outputs in life and increase employability, as digital literacy is considered an essential skill in today's digital world (Chetty et al., 2018). However, the concept of digital literacy is not merely practical, but is also an important term within media science. Digital literacy contributes to a greater understanding of users' ability to use and understand digital media in productive ways (Shen et al., 2019). In the sense of deepfake detection, digital literacy enables users

to critically evaluate the source, the content, and the intention, which are important cues when recognizing deepfakes.

Digital literacy has been a key area of focus in the Norwegian education curriculum since 1996 (Erstad 2006, p. 416). The emphasis lies mainly on skills directly linked to using the technology and includes evaluating sources and using ICT collaboratively critically. Perhaps because of this engagement, international research shows that compared to other European countries, Norwegians are on top when it comes to digital skills and that the majority of Norwegian pupils are highly digitally literate (Kunnskapsdepartementet, 2017; Fjørtoft, 2017).

A possible consequence of lacking digitally literate citizens in a population is digital divides. Tsai et al. (2017) argued that there are two levels of a digital divide. The first level involves having access to the technology in question, and the second level involves the need for skills and efficacy. The lack of these skills might affect how much citizens might benefit from using the internet. For instance, the possibility to access health-related portals, banking services, and other websites (Tsai et al., 2017). These digital divides might contribute to a growing distrust in the population among those who are not digitally literate and might lead to increasing knowledge gaps (Westerlund, 2019).

Therefore, being digitally literate is very important because knowing how to decide what is relevant and valuable and how to derive meaning while using technological devices is equally as important as using the technology itself (Chetty et al., 2018). Researching people's digital literacy in the context of deepfakes might help to understand the extent being digitally literate contribute to Norwegian's ability to distinguish real content from fake. Further on, I will examine how humans treat and evaluate media content and to which extent they might be flawed when deciding what content to believe.

3.3 Human selectivity and confirmation bias

Although confirmation bias is a term that originates from psychology, it has become increasingly important within media studies to understand how humans perceive and interpret media content, which in turn is highly relevant when researching human detection of deepfakes (Pearson and Knobloch-Westerwick, 2019; van der Meer and Hameleers, 2020). However, since the human detection of deepfakes is a relatively new research topic, previous research directly referring to this is scarce.

Even so, there has been extensive research on how humans subconsciously treat different media content and to which extent they manage to identify dis- and misinformation online, which in turn can contribute to a better understanding of how they also will treat content such as deepfakes. Both misinformation and disinformation are highly relevant terms when researching deepfakes. In this sub-chapter, I will look at what previous research has shown when it comes to being exposed to different forms of media, which kind of content is shown to be most effective, and how confirmation bias contributes to human's perception of media content.

From previous research, we know that images have a more substantial persuasive power than text and that humans remember what they have seen better than what they have read or heard on a general note (Vaccari and Chadwick 2020, p. 2). Simultaneously, humans are more likely to see audio and visuals as a more accurate description of the real world than textual descriptions. Because deepfakes usually are a mixture of both visual media and audio, it makes for quite a good way of spreading mis- and disinformation.

Psychologists have also found that people have confirmation bias (Schwebs and Østbye 2017, p. 214-215). This includes that we tend to remember better arguments that support our view, regardless of who promoted them, we usually interpret an argument to fit our vision of the world, regardless of how the argument was initially put forth. Additionally, we tend to expose ourselves to content that conforms to our point of view and avoid exposing ourselves to content that we disagree with (Pearson and Knobloch-Westerwick 2019, p. 467). Additionally, confirmation bias might increase social and political polarization (Sunstein, 2007).

Confirmation bias is perhaps especially important when looking at deepfakes because the theory shows that human's memory, perception, and exposure is deceitful (Sleegers et al., 2019). This, as well as the "truthiness effect" that explains how people tends to accept media's message if the content seems familiar, contributes to quite a few pitfalls in detecting deepfakes in real life (Newman et al., 2015; Brinsky 2015, p. 247). Deepfakes are often videos portraying celebrities, politicians, or other public figures, which contributes to the content more likely being accepted because of familiarity. Furthermore, because of the increasingly sophisticated technology, the sheer quality of the videos makes it harder for people to falsify only by what they see and hear (Vaccari and Chadwick, 2020).

Another challenge faced when it comes to deepfakes and the spread of disinformation is that attempts to correct false information are often useless. This is because of the so-called "illusory truth effect," which explains how continuous exposure to fake media, even if presented with the wish to correct that false information, increases the chances of fake media being remembered as true (Franks and Waldman 2019, p. 895; Aumyo and Barber, 2021). Because of this effect, fact-checking services such as Faktisk.no in Norway might actually be contributing to maintaining the spread false information and increase knowledge gaps (Kalsnes, 2019). Additionally, people tend to believe false headlines to a greater extent if they encounter them several times.

These are important considerations when it comes to researching deepfakes and their potential harm, considering that because of these biases, the spread of deepfakes might contribute to a widespread belief in the contents of a deepfake and might make it harder to convince the public of its falsehood. Because of confirmation bias and people's tendency to accept a message if it seems familiar, we might expect participants who have previous knowledge of the video or subject in the video are more prone to believe the content is authentic. In the next section, I will examine what previous research on the detection of fake media has shown, which can be used to set expectations for the results of this research and has formed the development of the survey.

3.4 Previous research on human detection and misinformation

Humans' ability to identify manipulated media have been the focus of research for some time, as the amount of such media has been increasing over the last decade (Nightingale et al., 2017; Holmes et al., 2016; Fan et al., 2012). As the quality of the fake media increases, human abilities to distinct them from real media decreases. Furthermore, a recent study suggested that people tend to overestimate their abilities to make distinctions between factual and fictional content (Thaler, 2019).

In Reuters Digital News Report in 2018, 47% of the Norwegian sample reported that they are quite or very sure that they would detect a fake news article (Newman et al., 2018). Simultaneously, 40% were quite or very uncertain of their ability to detect such fake content, and 14% responded that they do not know. This stood out compared to results

from other European countries and the United States, where the percentage of respondents being quite or very certain of their abilities to detect fake content lay between 71-84%. This gives reason to believe that a Norwegian sample might give different results than previous research done on other populations and supports the importance of researching this context.

A user study looking at humans' ability to detect digitally forged images showed a negative correlation between performance and age, and a positive correlation between performance and previous experience with digital images (Schetinger et al. 2017, p. 150). In their research, education had no significant impact on performance.

Further on, when researching human detection of fake face images that were extracted from an algorithm, Rössler et al. (2018, p. 13) found, using a quantitative study, that the accuracy of human detection could be equal to having a random guess when evaluating highly compressed images. An earlier study completed in 2012 researched humans' performance in recognizing fake face images generated by CGI (computer-generated imagery). Their results showed that detection accuracy varied with resolution and image compression but was, generally, more accurate than randomly guessing (Farid and Bravo 2012, p. 234).

Likewise, Nightingale, Kimberley, and Watson conducted in 2017 a study that showed that human's general capability to detect manipulated images are greater than chance (66%). However, although respondents did identify an image correctly, they often mislocated the manipulation location (Nightingale et al. 2017, p. 6).

In another study conducted in 2012, respondents were asked to differentiate between real photographs and CGI (Fan et al. 2012, p. 3). The results showed that the cues most important when deciding were eyes, skin, and lighting. Interestingly, the study also showed that ethnicity plays a significant role and that people's sensitivity is higher when it comes to faces of their own race. Additionally, they saw that details in the skin texture and glossiness of the skin play a significant role.

Previous research looking at the detection of actual deepfakes is quite scarce. Most research about human detection looks at manipulated images of different kinds, like those mentioned above. However, a quantitative study conducted by Khodabakhsh, Ramachandra, and Busch (2019) looked at six different kinds of deepfakes and peoples' accuracy in detecting them using an online survey. Their results showed a positive correlation between expected performance and the number of correct responses and a moderate positive correlation between age and the number of incorrect responses.

Furthermore, they saw a moderately negative relationship between age and the number of "uncertain" answers. However, the number of "uncertain" answers never exceeded 25% on any of the videos, showing that respondents were quite confident in their answers. Simultaneously, the most used clues to determine whether a video was real or fake were in the head and face area (Khodabakhsh et al. 2019, p. 5). However, this study only had 30 participants but remained an inspiration for the research conducted in this project.

Similarly, Korshunov and Marcel (2020, p. 5) conducted a study examining humans' ability to detect deepfakes and found that participants had a low number of uncertain answers but that the most sophisticated kinds of deepfakes fooled participants in 75.5% of the cases. This study had approximately 20 participants and compared human detection capabilities with two deepfake detection algorithms, which showed that humans perform quite poorly.

The introduced concepts and previous research will provide this thesis with a sound theoretical framework and will allow for a broad description of the research results.

3.5 Summary and presentation of hypotheses

Based on the previous research concerning human detection and misinformation, we might expect from our research to see a positive correlation between both age and digital literacy and number of correctly identified videos.

Education is a point of discussion included in several previous research projects because of its close connection to higher socioeconomic status. Nevertheless, education has not shown to significantly impact human capabilities to recognize fake media (Khodabakhsh et al., 2019; Schetinger et al., 2017). This is interesting because students of higher education are taught digital literacy through the critical evaluation of information and sources, which in turn prevents the development of knowledge gaps. Education is also considered a clear factor of socioeconomic status, which is one of the key factors when explaining knowledge gaps in a society. Moreover, digital education is considered one of the most important measures to prevent the spread of misinformation. Because of this, and because higher education seeks to teach students critical thinking in general, a hypothesis looking at education was still included.

Digital education is thought to be an essential tool to combat the spread of fake content online (Vaccari and Chadwick, 2020; Diakopoulos and Johnson, 2020). This is because deepfake-detection technology is still flawed, and scientists are struggling with making new software programs that can detect deepfakes with 100% accuracy. Hence, people's digital literacy is considered an essential factor when distinguishing real content from fake.

Even though trust in democracy and governance in Norway is quite high, trust in news is relatively low, despite relatively low political and social polarization levels (Newman et al., 2018). Research showed that this lack of trust is related to politics, with far-right voters having decreased trust in the news. However, as previously mentioned, concerns about the exposure of fake content are notably lower in Norway than in other countries. Simultaneously, fear of twisted stories to push an agenda is considered more concerning than entirely made-up stories. Newman et al. (2018) added that this might be due to a strong tradition of objective reporting in Norway. This tradition of objective reporting and low concerns might make Norwegians vulnerable to fake content, which might contribute to growing knowledge gaps. Because of this, deepfakes and misinformation might be a bigger problem for Norway than it is in other countries, because it takes advantage of their low concerns and tradition for objective reporting. Hence, a hypothesis regarding participants trust in news outlets was included in the survey.

In previous research, previous knowledge about the subject in the video decreased the number of "uncertain" answers (Khodabakhsh et al., 2019). This shows that in their sample, prior knowledge contributes to more certainty when deciding. We might expect that participants with previous knowledge of the subjects perform better when classifying the video and are less likely to give an uncertain response.

Although some research on the detection of fake media has shown no difference in gender when classifying fake content (Farid and Bravo, 2012; Fan et al., 2012; Khodabakhsh et al., 2019), this will still be included as a control variable to confirm previous results. Moreover, when Khodabakhsh et al. in 2019 researched peoples ability to detect deepfakes

on a small sample, their results showed that the level of expected expertise had a positive correlation with the number of correct answers. With these results in mind, there is reason to believe that expected performance might affect actual performance.

As previously noted, deepfakes are most commonly made of politicians or other celebrities because there are usually a high number of images available on the internet (Spivak, 2019). Similarly, previous research has shown that having previous knowledge of the subject portrayed in the video will decrease the number of uncertain answers (Khodabakhsh et al., 2019). Because of this, political interest might affect participant's performance because it increases the probability that a participant will have previous knowledge if the deepfake portrays a politician.

Further, a question included in the survey referred to the number of hours spent on the internet per day. An increased number of hours spent on the internet increases the probability of having encountered a great deal of digital content and it increases digital literacy, which, in turn, reduces knowledge gaps. Moreover, news outlets and social media use manipulated stories and media to make clickbait headlines to increase their revenues, and deepfakes might become a more prevalent part of this in the future (Aldwairi and Alwahedi, 2018). Correspondingly, previous research showed that more exposure to digital content positively influenced people's ability to detect fake digital content (Schetinger et al. 2017, p. 147).

Lastly, when researching human's abilities to detect fake media content, researchers sometimes include a question asking what cues were important when making their decision (Khodabakhsh et al., 2019; Fan et al., 2012). This is an important question for several reasons. For instance, it is important to look at properties of the videos and not just properties of the respondents. Further, it gives insight into the thought-process when participants evaluate media content and whether their cues are based on real shortcomings of the media content or if participants are trying to find mistakes that are not there.

Based on this, I present the following hypotheses that will be tested using data from a survey developed for the purpose of this study:

> *H1: Older respondents will perform worse when classifying deepfakes than younger respondents.*
>
> *H2: Respondents with higher education will perform better when classifying deepfakes than respondents without higher education.*
>
> *H3: Respondents who report high digital literacy will perform better when classifying deepfakes than not digitally literate respondents.*
>
> *H4: There is a negative relationship between having increased trust in news outlets and the ability to identify deepfakes correctly.*
>
> *H5: There is a negative correlation between previous knowledge of the subject in the video and the number of uncertain responses.*

Additionally, control variables were added to better understand the factors affecting Norwegian's ability to detect deepfakes. These control variables include gender, expected

performance, political interest, internet use per day, previous knowledge of the video in question, and which cues were used to decide.

4 Method

This chapter will explain why the specific method was chosen, the sample size and distribution, and explain how the survey was developed. Further on, I will unfold the process of analyzing the data, discussing the quality of the research, and research ethics.

4.1 Choice of method

The chosen method to answer the research question was a quantitative digital survey made on the platform nettskjema.no, created by the University of Oslo. A survey was thought to be the best design for answering the research question because it allows for quantification of large amounts of data, which in turn can provide statistical answers to the extent to which Norwegians can correctly identify deepfakes (Evnas and Mathus, 2005; Mullinix et al., 2015). Because of this property, surveys of different forms have also been the preferred methodical design in previous research on human detection of fake media (Khodabakhsh et al., 2019; Schetinger et al., 2017; Vaccari and Chadwick, 2020). Making a digital survey allows for collecting a considerable number of respondents in different age groups and with different educational backgrounds. This is of great significance for the research's ability to give a statistical description of the population. Simultaneously, using an online survey is more beneficial because all respondents were presented with the same questions, the responses were anonymous, and because it is online, it might lower the threshold for people to take part (Ringdal 2013, p. 190). The survey also allowed me to get respondent's honest answers when identifying the videos, without having the social pressure that might have occurred if the research was carried out face to face (Evnas and Mathus, 2005).

Because there was no public dataset from previous identification of deepfakes from a Norwegian sample, I had to generate it for this survey. Since I depended on the respondent's ability to play a video, preferably with sound, performing the survey online was considered the best way. This assured that anyone completing the survey had the digital literacy to navigate the web since they had been able to find and enter the survey. The service Nettskjema.no was chosen as the provider of the online survey setup because NTNU has signed a data processor agreement with the University of Oslo, which reassures that the platform is safe for research use and that respondent's privacy is preserved. Although no sensitive personal data was gathered from participants, it is still crucial that respondent's responses are safely handled.

The research design is a cross-sectional study, meaning that observations were only made at one point in time, intending to compare the participants and look for potential variation and co-variation (Skog 2017, p. 71). This survey was open for responses from February 22nd to March 30th, 2021. The participants were chosen with the goal of all age groups being represented. Because the online survey was distributed through advertising on Facebook, segregation based on age was effortless and more effective than surveying using paper. Previous research has also shown that properly conducting an online survey and recruiting participants using Facebook create datasets that to a large extent are as representative and diverse as other forms of survey data (Bhutta, 2012; Schneider and Harknett, 2019; Mullinix et al., 2015; Evnas and Mathus, 2005).

The general problem of such cross-sectional studies is that there is a higher chance of the correlations being spurious, as the potential co-variation might be based on other variables than those in the survey. The reason why some participants have high values on an independent variable, as well as the dependent variable, is not necessarily because the independent variable is a causal factor of the dependent variable (Skog, 2017). However, this will be further evaluated when discussing linear and logistic regression assumptions in the analysis.

4.2 Structuring the survey

In this sub-chapter, I will describe the development of the survey. Developing and working with the survey was an essential and time-consuming part of the research, and each part of the survey was well thought through with potential strengths and weaknesses in mind.

The structure of the survey can be divided into four parts. The first part included the collection of consent and demographic data. The second was an introduction to the term deepfakes, what it means, and what it includes, and a question about how well the participant considered their performance (see attachment 1). Then came the central part of the survey where participants were presented with the 16 videos and questions about the video's authenticity, previous knowledge of the video and subject, and which clues affected their decision. The final part was a feedback page, where participants could write feedback on the survey or the topic.

The survey was inspired by several previous research designs looking at detecting of fake media (Khodabakhsh et al., 2019; Fan et al., 2012; Nightingale et al., 2017; Korshunov and Marcel, 2020). However, because the survey for this research was set to target the Norwegian population and to intend to attain an increased number of observations, the survey was somewhat different from the surveys mentioned above.

When considering that the Norwegian population is standing out from other European countries, both because of their high level of digital literacy but also because of their low exposure to fake media content, it was important to make sure the respondents were well-informed on the topic of deepfakes, and on the task they were about to carry out (Fjørtoft, 2017; Newman et al., 2018). Hence, they were introduced to what deepfakes are with a definition and a short explanation of which kinds of deepfakes exist. However, only deepfakes of the most sophisticated kind would be used in the following survey. Further, respondents could read an explanation about how the survey would go about, and that there were 16 videos approximately 10 seconds long, and that half of the videos were genuine, and half were deepfakes. They could watch the videos numerous times and on full screen. They were also recommended to complete the survey on a computer or a tablet. The definition of deepfakes was included to ensure that respondents were fully aware of what they were looking for since they might not have been exposed to deepfakes previously. When respondents were asked about digital literacy, the definition of the term was also included to ensure that respondents knew which abilities they should consider. Although this might have bettered participants' understanding of the term, the fact that it is still a question of subjective perception will be discussed.

When deciding the number of videos to include in the survey, some important factors were considered. A high number of videos would also increase the length of the survey,

demanding respondents to set aside an increased amount of time. Conversely, a low number of videos would decrease the quality of the research and decrease the chances of obtaining statistically significant results in regression analysis. Using responses from pre-testing, the final number of videos became 16, which lead to the expected response time being approximately 20 minutes. The 16 videos included 8 deepfakes portraying Russian politician Aleksej Navalnyj, the late actor River Phoenix, Queen Elizabeth II, the actor Steve Carell, the artist Dua Lipa, the actress Margot Robbie, South Korean dictator Kim Jung-Un and Danish prime minister Mette Frederiksen. The other 8 videos were authentic and portraying German Chancellor Angela Merkel, British Prime Minister Boris Johnson, actress and artist Zendaya, Swedish union leader Stefan Löfven, actor Jared Leto, American whistleblower Edward Snowden, previous first lady Melania Trump, and actor Nicolas Cage. The videos lasted approximately 10 seconds, as this was considered enough time to get an impression of the video, but not enough time to necessarily perceive the context or content of the video. Hence, evaluating their authenticity would solemnly be based on looks and audio instead of content. In previous research considering deepfake detection, video lengths have also varied from 4 to 11 seconds (Khodabakhsh et al. 2019, p. 2; Tolosana et al. 2020, p. 135).

Because the total number of videos was set to 16, I decided only to include deepfakes of the most sophisticated kind, made with GAN technology or similar. The choice to only include deepfakes made by GAN technology was based on the aim to make the survey as realistic as possible in terms of how the respondents might come across deepfake videos in real life. Although the YouTube videos rarely stated whether GAN technology specifically was used, the sophistication of the video and the overall looks was used to evaluate the quality. As mentioned in the introduction, in real-life instances where deepfakes have been spread, they have been sophisticated enough to fool many people. When discussing the realistic feeling of the online survey, the apparent difference between being exposed to deepfakes in real life and completing this survey is that participants are taking part in a survey where the main objective is to identify deepfakes. Hence, they will most likely be more concentrated on finding cues to identify deepfakes than they will be in real life. However, by using the same kind of video that is expected to be used in misinformation, a possible outcome could be to raise awareness, hopefully leading to participants being more critical when evaluating content in the future.

Other factors had to be considered when selecting the videos to include in the survey. Although the internet is full of sophisticated deepfakes, many of them are deepfakes made from scenes in movies. A criterion when selecting the videos was to limit the number of videos from movie scenes, enabling the audience to recognize a deepfake based on previous knowledge of the movie instead of the looks of the video. Hence, videos of politicians were favorable because the respondents would answer based on what they see instead of previous knowledge of the video. Videos of politicians also have the advantage that most politicians conduct interviews and press conferences regularly, which decreases the chances of respondents recognizing the exact video from having seen it before. Moreover, deepfakes with politicians might be the kinds of deepfakes we as citizens might interact with the most in the future on social media (Donovan and Paris, 2019).

Hence, only three videos from movies or tv-shows were included in the survey, two deepfakes and one real. Further, 8 of the videos were of politicians from the United States, North Korea, Denmark, Sweden, and Britain. Four deepfakes and four authentic. The last five videos portrayed celebrities in a perfume commercial, interviews, and one playing

saxophone. All videos included in the survey were of the same quality, 480 pixels. Having other faces in the video was considered a potential distraction and was kept at a minimum to decrease disruption. Further on, on a general basis, I tried to include videos of the same kind, whether it be politicians' press conferences or speeches, celebrity interviews, or commercials. These are also some of the most common forms of deepfake videos on YouTube, except for movie scenes. By including deepfakes and real videos of the same kind, I was limiting the respondent's possibility of identifying deepfakes based on their typical form instead of their looks.

The survey went through a pretest where a first draft of the survey with 31 videos was distributed to five classmates to get feedback. The feedback was thorough and detailed and led to quite a few changes in the survey. Additional options were added to the question about the factors that lead to the respondent's decision and interest in politics. The order of the questions gathering demographic information was also changed to get a more natural structure. Additionally, a question was added on the feedback page asking if the respondents have any other thoughts about deepfakes that they would like to share. This question was added because the theme of the survey might be new and possibly a bit scary and might have left respondents with a sense of being anxious or frightened. Hence, having a place to vent their thoughts might have been valuable to them.

The survey was open from February 22nd to March 30th, 2021. The respondents were recruited using Facebook, LinkedIn and the online bulletin board of NTNU. A Facebook page was created to publish the link to the survey and post the correct answers when it was closed for further responses. When the post containing the survey was published, I used Facebook ads to boost the post and advertise it on users' walls to increase engagement. When boosting a post like that, Facebook allows you to target your audience with great precision, making it effortless to reach specific target audience groups if they are underrepresented at a given time.

4.3 Variables

This sub-chapter will address the dependent, independent, and control variables used to answer the research question. To give a clearer overview and reduce confusion, the sub-chapter is divided into variables used in the linear regression and variables used in the logistic regression. The two measures different parts of the research questions and contributes to the research in different ways. The variables will be explained in terms of measurement level and how they were measured in the survey. This will contribute to greater transparency and increase the replicability of the survey.

4.4 Variables used in the linear regression

The linear regression was used to measure the correlation of the demographic information on the ability to correctly detect deepfakes, and the level of uncertainty.

4.4.1 Dependent variable

The first dependent variable in the linear regression concerns the total number of correct responses for each respondent. The variable was measured in the survey by the question "Is the current video of *name* real or fake?" with three alternatives being "real and authentic," "unsure/do not know," and "fake." These variables are at a nominal level, as the answers are mutually exclusive (Ringdal 2013, p. 90).

In the coding book, the correct answer was coded to 1, the incorrect answer was coded to 2, and an unsure answer was coded to 3. However, upon analysis, these variables were coded to be dummy variables, where 0 equaled an incorrect or uncertain answer (2 or 3 in the codebook), and 1 is a correct answer. Further, the variable used in the analysis was generated by adding all these variables together, adding the occurrences of 1, and generating a new continuous variable with 17 categories going from 0-16. The variable was named "Classification of deepfakes." The variable had 682 observations with a mean of 9.17 and a standard deviation of 2.5.

This variable can examine the part of the research question concerning the overall ability amongst respondents to detect deepfakes. Figure 1 shows the distribution of correctly identified videos.

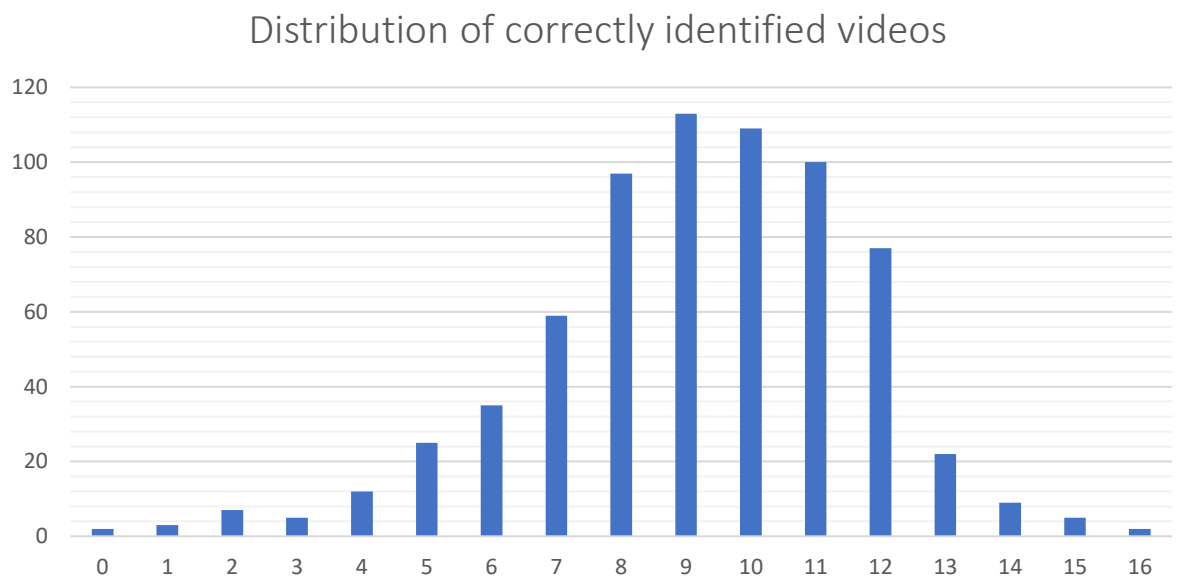


Figure 1: Distribution of correctly identified videos.

The second dependent variable examined the number of uncertain answers given by respondents and is called "Uncertainty when classifying." This variable was interesting to include in the analysis because it gave insight into the respondents' certainty and confidence when answering and allowed me to research whether different factors influence the number of correct answers and the number of uncertain ones. Including an "uncertain/do not know" category when classifying videos also allows me to pick up on respondents who most likely would have randomly guessed if the option was unavailable, adding another dimension to the research. The Uncertainty when classifying variable is made up of 16 variables evaluating the authenticity of the videos. This is based on the same variable as used in the first dependent variable and is on a nominal level. However, when creating this variable, the incorrect and correct answers were coded to be 0, and the uncertain answers were coded to be 1. For the analysis, the variable concerning the number of uncertain answers was generated by adding together all these dummy variables, creating a new variable on ratio level with 17 categories ranging from 0-16. This variable also had 682 observations, a mean of 3.77, and a standard deviation of 2.77.

This variable allows for examination of respondents' confidence when classifying videos, since it measures the number of uncertain responses. Respondents with a high number of

uncertain answers could be thought to have low levels of confidence in their judgement. Figure 2 shows the distribution of Uncertainty when classifying.

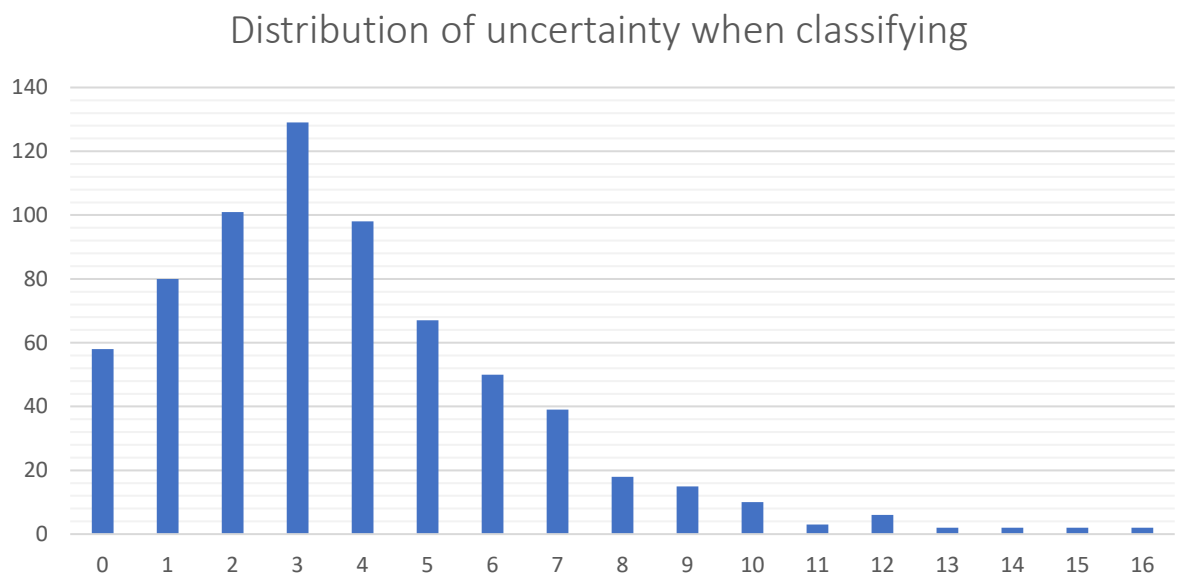


Figure 2: Distribution of uncertainty when classifying.

As Figure 2 shows, the distribution of Uncertainty when classifying has a noticeable shift to the left, compared to the variable Classification of deepfakes. Most respondents gave a small number of uncertain answers.

4.4.2 Independent variables

"Age" is a variable on a ratio level as the distance between the age groups is equal, there is a rank order between the age groups, and there is a true zero value (Ringdal 2013, p. 90). There was a total of 12 age groups that handled every age from 18 to >70 years old, with each age group containing approximately 5 years, except the age >70. Upon analysis, the variable was recoded, so the age groups ranged from 1-12 in ascending order. This variable will answer the hypothesis concerning the positive correlation of age on the ability to detect deepfakes.

"Education" is a variable on an ordinal level since there is no definite difference between the categories, and there is no relevant ratio between the categories (Ringdal 2013, p. 90). The variable was measured in the survey by asking the respondent's highest level of completed education, with categories being primary and secondary school, high school, higher education - no completed degree, higher education - one year program, higher education - bachelor's degree, higher education - master's degree and higher education - Ph.D. The categories were coded from 1 to 7 in the order they were mentioned above in the analysis and will contribute to answer the hypothesis concerning the influence of education on the overall ability to detect deepfakes.

"Digital literacy" is also an ordinal variable where respondents were asked to evaluate their general digital literacy after being given a definition of the term. The categories were no digital literacy, low digital literacy, moderate digital literacy, quite high digital literacy, and very high digital literacy. The category was coded from 1 to 5 in the order mentioned here.

This variable will allow for examination of the hypothesis stating that there is a positive correlation between the level of digital literacy and ability to detect deepfakes.

The last independent variable of the linear regression is "Trust in news," which is also on an ordinal measurement level and was measured in the survey by asking respondents, "How high is your trust in Norwegian news media outlets?". Categories included very low trust, low trust, moderate trust, high trust, and very high trust, coded from 1-5 in ascending order. This variable will answer the hypothesis concerning the negative correlation between trust in news and the ability to detect deepfakes.

4.4.3 Control variables

In addition to the independent variables mentioned above, a few control variables were added to better understand how other factors might influence the ability to detect deepfakes. These control variables are gender, expected performance, political interest, and internet use.

Gender is a nominal variable. It was measured by asking respondents which gender they identify as, with alternatives being female, male, and other/do not wish to enter. In the codebook, 0 was female, 1 was male, and 2 was other/do not wish to enter. However, in the analysis, the variable was recoded to a dummy variable where 0 meant female, 1 meant male, and the 6 participants answering "other" was dropped from the sample. This new variable was called "Man."

"Expected performance" is an ordinal variable that was measured by asking, "How well do you think you will perform when identifying deepfakes?" with the categories very poorly, poorly, moderately, good, and very good. These were coded to number codes 1-5 ascending order from worst to highest.

"Political interest" was measured by asking respondents how politically interest they consider themselves to be. Categories were "very uninterested," "quite uninterested," "a bit uninterested," "neither nor," "a bit interested," "quite interested," and "very interested." Political interest is an ordinal variable, and categories were coded from 0-6 in ascending order from least to most politically interested.

"Internet use" was measured by asking respondents how many hours they spend on the internet on a typical day, with categories being 1-2 hours, 3-5 hours, 6-8 hours, 9-11 hours, 12-15 hours, and more than 15 hours. This is a ratio leveled variable where categories have a true zero value, there is a rational rank between the groups, and the difference between the groups is constant (Ringdal 2013, p. 90). The categories were recoded to numbered values ranging from 1-6 in ascending order.

4.5 Variables used in the logistic regression

In addition to linear regression models, logistic regression models were also included to measure the influence of previous knowledge of the subject and the video. The difference between linear regression and logistic regression is that, in logistic regression, the dependent variable is a dummy with only two categories. Logistic regression measures the probability of the dependent variable having the value 1. In my case, this allows me to examine whether previous knowledge of the subject or the video increases or decreases the probability of giving a correct answer.

4.5.1 Dependent variable

The dependent variables used in the logistic regression models are dummy variables made from the question "Is the current video of *name* real or fake?". For each video, the incorrect answer and an "unsure" answer was coded to be 0, whereas the correct answer was coded to be 1. In this way, I can look at how previous knowledge of each subject and video influences their performance separately. The variables were named after the subject portrayed in the given video.

4.5.2 Independent variables

"Previous knowledge of subject" is the only independent variable of the logistic regression and is a variable on the nominal level because the number codes are only labels describing which group the participants are located (Ringdal, 2013). This was measured in the survey by asking the respondents whether they had previous knowledge of the subject following the introduction of a new video. Categories included "yes," "unsure," and "no." Upon analysis, these variables were dummy coded so that 0 included "unsure" and "no" responses, and 1 included "yes." This variable will answer the research question stating that there is a negative correlation between previous knowledge of the subject and the number of uncertain responses.

4.5.3 Control variables

"Previous knowledge of video" is a control variable of the logistic regression model and was measured by asking respondents whether they had seen the video or other versions of the video previously. This is a nominal variable with mutually exclusive categories: yes, no, and unsure. Upon analysis, the variable was recoded to a dummy variable where no and unsure answers were recoded to 0 and yes being recoded to 1.

Lastly, "Relevant cues" was a question concerning the different cues that respondents used when evaluating the authenticity of a video. This variable was not analyzed in any regression models but examined on its own. It was measured by asking respondents what an essential factor for their decision was. Categories mentioned were previous knowledge, head/face, eyes/blinking, background, body, shadows, movements, words/speech, sound/voice, synchronicity in movements, synchronicity in sound/voice, unsure, and other. Respondents could choose as many categories as relevant. Upon analysis, the variable was divided into cues used to evaluate deepfakes and real videos, where the occurrence of each cue was measured and presented. This variable was not included in any regression analysis but was analyzed alone.

4.6 Sample

Table 1 shows how the sample is assembled, with the average value of each variable, standard deviation, and the said variable's minimum and maximum value. Because variables were recoded to have numerical values, they are represented by numerical groups instead of the actual value.

Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Gender	682	.512	.551	0	3
Age	682	6.334	2.882	1	12
Education	682	4.677	1.519	1	7
Digital literacy	682	3.823	.756	2	5
Internet use	682	2.452	.979	1	6
Trust in news	682	3.532	.798	1	5
Interest in politics	682	4.4	1.433	0	6
Expected performance	682	3.236	.736	1	5

Table 1: Descriptive statistics of the sample

The survey consisted of 682 respondents, where approximately 48,5% were men, 50,6% were women, and 0,9% did not wish to specify. The average age group of the respondents was 41-45 years old, and the average education level was a bachelor's degree. On average, respondents considered their digital literacy quite high, and the average participant reported that they spent approximately 3-5 hours on the internet. Respondents had, on average, high trust in news media outlets and were slightly interested in politics. Further on, the average expected performance amongst respondents was moderate.

4.7 Analysis

Two linear models of multiple regression were included in the analysis to give proper insight into whether the independent variables influenced Norwegian's ability to recognize a deepfake and their confidence when responding. Linear regression allows for statistical analysis of whether there is a linear correlation between a change in an independent variable and a change in the dependent variable (Lewis-Beck 1993, p. 1). Another two logistic regression models were included to analyze whether previous knowledge of either the video or the subject influences the number of correct answers or the number of uncertain answers.

The linear regressions concern the question which factors affect Norwegian's ability to detect deepfakes. The linear regression was divided into two. The simple model was a basic model of the dependent and independent variables before presenting an improved model where control variables were included, and variables were correctly specified to better explain Classification of deepfakes. In the simple model, I used Classification of deepfakes as a dependent variable and Age, Higher education, Digital literacy, Trust in news, and Previous knowledge of subject as independent variables. This gave insight into whether any independent variables showed a statistically significant linear correlation with Classification of deepfakes. In the adjusted models, I added the control variables Male, Expected Performance, Political interest, and Internet use and correctly specified the variables from the first model. This resulted in a more fitting description of the effect of the independent variables on Classification of deepfakes and gave insight into how different categories of independent variables have different effects.

Another linear regression was added with Uncertainty when classifying as the dependent variable. In the same way as with the first linear regression model, a simple regression model was first presented using only the independent variables. Then an improved model containing control variables and correctly specified variables was presented, striving to find

the model most suited for explaining Uncertainty when classifying. These models were included in the analysis to give a broader insight into participants' general confidence when evaluating videos and which factors affected this confidence.

Further, two logistic regression models were included to evaluate the effect of having previous knowledge of the video in question, or the subject in the video, on correctly identifying the video in question or giving an uncertain answer. A logistic regression model differs from linear regression because it calculates the odds of being in one category versus the other (Lewis-Beck, 1993; Stoltzfus, 2011; Meurer and Tolles, 2017). These logistic models were included in the analysis because it gives insight to whether Previous knowledge of video or Previous knowledge of subject influences the odds of correctly identifying a video or giving an uncertain answer. These allow me to further answer what influences Norwegian's ability to detect deepfakes and the hypothesis concerning whether Previous knowledge of the subject increases respondent's confidence, leading to lower odds of giving an uncertain answer.

Finally, an analysis of different categories within "Relevant cues" was completed to give a broader understanding of what cues respondents use when evaluating videos and whether some cues are used more than others, looking at real videos and deepfakes separately.

4.8 Quality of research

When discussing the quality of the research, we need to look at validity, potential measurement errors, generalizability, research ethics, and challenges when collecting data that might have impacted the results. Some of the limitations will also be examined in the discussion.

To evaluate the research's validity, we need to examine potential measurement errors and representation errors. One challenge with surveys like these is that I leave my complete trust in people's self-reporting, which might substantially affect the validity and reliability (Scharkow, 2016). Especially when evaluating videos that were initially collected from Youtube.com, nothing stops the participants from looking up these videos on their own, while completing the survey. However, although it decreases the validity, it increases the realism of the project (Ringdal, 2013). The information that the videos were collected from YouTube was never disclosed to the participants, decreasing the chances of respondents looking them up. Another point that affects the reliability and validity is that when asking respondents about education, the alternatives only include higher education from a college or university, but not an alternative for people who have a trade certificate, journeyman's letter, or the equivalent. Hopefully, however, the respondents who completed the survey responded with the alternative that equals the number of years they spent educating themselves.

To further evaluate the data's validity, I need to evaluate whether the variables included in the data will measure what needs to be measured to answer the research question. The research question is "To what extent are Norwegians able to recognize deepfakes and which factors affect this?". The first part of the research question regarding the extent to which Norwegians can recognize deepfakes is measured by the questions concerning the authenticity of each video, where participants are either correct, uncertain, or incorrect. Making a variable out of all 16 such answers, I would argue, is an accurate measure of the extent to which Norwegians can recognize deepfakes.

However, the second part of the question concerning which factors affect their ability to detect deepfakes needs evaluation. On the one hand, variables that from previous research have shown to be relevant have been added to the models, some of which showing significant effect. However, like previously mentioned, the nature of cross-sectional studies includes that we can never know with complete certainty that the correlations are spurious or causal. Nevertheless, considering the heavy amount of research conducted on the detection of fake images, deepfakes, and misinformation, there is reason to believe that the correlations are is not only spurious.

Using Facebook to distribute and recruit participants for an online survey has been a theme of discussion within the sociological methodology. This is because Facebook is a social platform on the internet, which breaks traditional norms of data collection (Groves, 2011). Since both the survey and Facebook are online, one problem is that subjects in the target audience who lack the needed computer skills and/or equipment will not participate (Couper et al., 2007). Hence, only citizens of a particular socioeconomic status could participate. This is called under-coverage and describes the inability of the survey to select some groups of the target population. This is a form of selection bias that will limit the generalizability of the data since the survey's target audience also includes groups that do not have access to the internet (Bethlehem, 2010). Another form of selection bias is self-selection, which means that it is up to each person to participate in the survey without the researcher having any control of the selection process. This goes against the principle of probability sampling, which means that the estimated results in the analysis could be biased (Bethlehem, 2010).

However, because internet access has increased and Facebook has grown to become one of the largest social platforms, the share of citizens who are not digitally literate and do not have a Facebook profile continues to shrink (Schneider and Harknett, 2019). Hence, Facebook has become a more functional and more representative tool to recruit participants for online surveys (Schneider and Harknett, 2019; Bhutta, 2012). Therefore, I will argue that the problem of under-coverage does not impose a substantial threat to the results of the analysis. In Norway, it is estimated that approximately 3,3 million citizens have a Facebook profile (Tankovska, 2020).

The advantages of using an online survey and distributing on Facebook are many. Firstly, web-based surveys are faster and more available, as it removes the need to manually handle the data collected and open to a much larger geographical audience than a paper-based survey would (Evnas and Mathus, 2005). Facebook also allows for detailed targeting for your ad based on their users' demographic data. Additionally, since the penetration of the internet rate continues to increase, samples from online surveys have also become increasingly representative (Bhutta, 2012). Moreover, advertising on Facebook is, in general, cheap. For 3200 Facebook users to click the URL in the post, I paid approximately 2000 NOK, meaning that one click cost approximately 1.6 NOK, although a large share did not complete the survey.

When distributing the survey, one potential flaw that might have contributed to the low completion rate was not describing the term deepfake in the Facebook post. This might have led to a skew in the distribution of participants because people who do not know what a deepfake is beforehand might be more reluctant to enter and complete the survey. This was pointed out in the comment section of Facebook, saying that they did not wish to participate without knowing what a deepfake was. This can also be seen by looking at the

reach and rate of engagement on the Facebook post. According to Facebook, 21400 users were reached by the advertising of the post, whereas 3200 users engaged by clicking the post, but only 682 completed the survey. Because of this, I will argue that a more welcoming title or using a picture of myself to give the audience a human face could have increased the number of participants (Bhutta, 2012; Schneider and Harknett, 2019)

Further on, when boosting a post on Facebook, there are several decisions you need to make to let Facebook know whom they should target. One of these decisions is your goal for the ad. Getting more website visitors was chosen as the goal so that more people would click the URL and complete the survey. However, this is another point where people in the target group might have been excluded if Facebook did not see them as people who are likely to press an URL. How Facebook measures which users are more likely to click an URL is a so-called black box and is unknown to researchers. This, as well as their power to change their algorithm without anyone knowing, is a limitation for the use of Facebook when advertising online surveys (Schneider and Harknett, 2019).

Another factor that could affect generalizability is the distribution of the sample completing the survey (Ringdal, 2013). If some age groups are marginalized or exaggerated, the results might not be generalizable. Because of this, I paid close attention to the age distribution of the respondents throughout the time the survey was public and changed the target audience of the Facebook ad according to which age groups were lacking respondents. Doing so increased the representativeness of the sample significantly. However, their age was the only feature being used to determine the target audience, as opposed to education or region of the country. This led to the misrepresentation of respondents from specific educational backgrounds. This influences the research's generalizability (Ringdal, 2013).

However, because the number of observations is relatively large, with 682 participants, this increases the chances of being able to generalize the results beyond the sample, with a relatively high representation of citizens above the age of 60, although they are underrepresented in the sample (SSB, 2021). This was the age group thought to be the hardest to recruit. Simultaneously, no poststratification weights were used on the data set, meaning that there is no compensation for underrepresented groups. To summarize, there are challenges and weaknesses in the research that will affect the quality. However, the study provides detailed insight into the Norwegian sample in the context of misinformation and deepfakes and will still be an important contribution in the research field.

4.9 Challenges when collecting the data

One of the biggest challenges when collecting data, particularly data from participants at the age of 60 and upwards, was that they would end the survey before completing it. Some explained this by saying that they could not evaluate a video without judging the content and context of what the subject was saying. This feedback was posted on the Facebook post and in the feedback-question in the survey. A thread of comments arose saying that they felt it was useless to identify fake videos in this way when they would do more extensive research to evaluate whether a video was fake or not in real life. This might have affected the research results by eliminating participants from that age group. This was the most significant challenge when collecting the data, and this could be a potential systemic measurement error. By communicating better that the purpose of the research was to see

people's ability to detect deepfakes solely from their looks and not their content, this problem might have been eliminated.

4.10 Research ethics

An ethical issue that I was faced with when making the survey was whether or not to include the solution to which videos were fake and which were real at the end of the survey, as other scientists have done in their research (Khodabakhsh et al., 2019; Schetinger et al., 2017). This was done to make the survey feel less like a competition or a test, and more of a regular questionnaire, so that the respondents would not be tempted to retake the survey to improve their "score". However, one can argue that it is unethical to leave the respondents not knowing which of the videos are fake and that I participate in spreading uncertainty and doubt amongst the respondents. When looking at feedback from respondents, both in the survey and on Facebook, the vast majority states a wish to see the results. For this reason, I could have included the results in the end to improve the satisfaction amongst participants.

However, because of my inability to control the respondent's participation without collecting significantly more intrusive sensitive personal information, I chose not to include the respondent's results after completing the survey. Simultaneously, to meet the respondents' expectations, I included in the survey a link to the Facebook group where the results would be posted once the survey was closed. Moreover, even though I collected the participant's informed consent and explained the structure of the survey, I might not have been meeting their expectations by not including the results.

Informed consent was collected from participants before collecting demographic data, and anyone with access to the link could participate. Since no sensitive personal data needed to be collected to complete the survey, the project did have to be accepted by the Norwegian Center for Research Data (NSD). Moreover, the data material was anonymized from the start because no personal data that could recognize respondents were collected. The data was stored locally, and a report of the results could be collected from nettskjema.no. Since the data was anonymized from the time of collection, I did not see a problem with keeping the data on my personal computer. When the research project has ended, estimated to be June 7th, the responses will be kept on an encrypted memory-stick, because the anonymized results might be used for further research.

5 Analysis

In this chapter, I will present the results of the analysis in an attempt to answer the research question: To what extent are Norwegians able to recognize deepfakes and which factors affect this? I will present the data collected from the online survey and perform regression analyses to examine the relationships between the dependent and the independent variables. The chapter will be divided into two parts. First, a presentation of the descriptive statistics and an overall review of the data material, then in the second part, presenting linear regression models, logistic regression model, assumptions of the quality of these models, and a review of the most used cues when categorizing videos. Together, these will give a greater understanding of Norwegians ability to identify deepfakes correctly and factors that influence this ability.

5.1 Overview of data material: Overall ability to detect deepfakes in the population

Figure 3 shows the overall distribution of correct, incorrect, and uncertain responses for each of the 16 videos. This shows which videos were most difficult to classify correctly and which videos respondents were most uncertain.

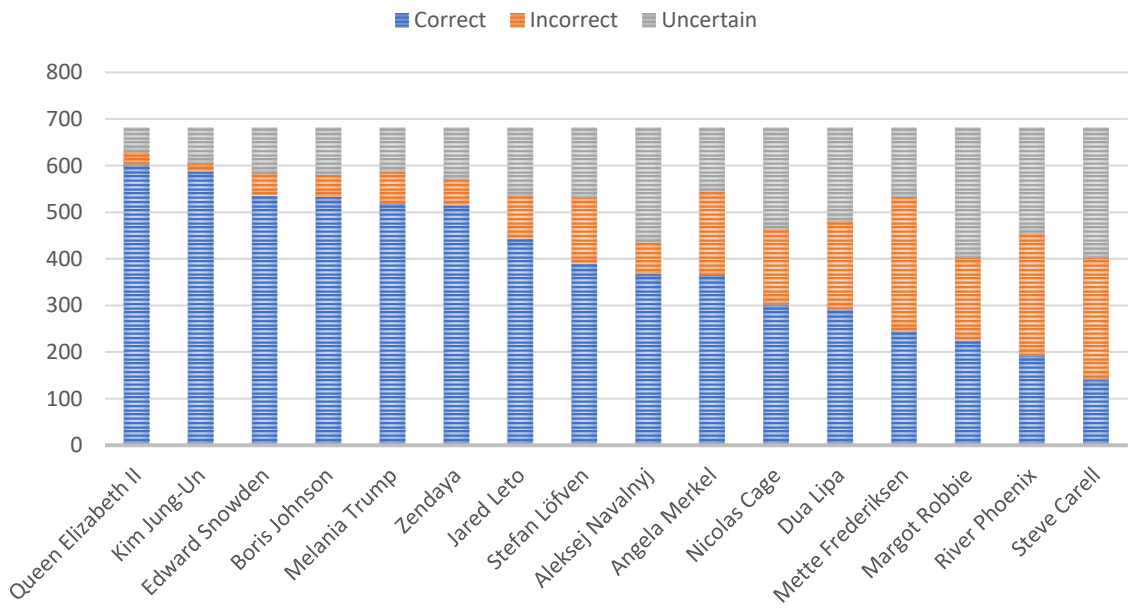


Figure 3: Distribution of correct, incorrect, and uncertain answers for each video.

As Figure 3 shows, the most often correctly identified video was the video of Queen Elizabeth II, with 88.3% correct responses. The second and third video with the highest share of correct responses was Kim Jung Un with 86.2% and Edward Snowden with 78.4%. The videos that had the lowest shares of correct responses were Margot Robbie with 33%, River Phoenix with 28.3%, and Steve Carell with 20.5%

The video that had the most "uncertain" responses was Steve Carell with 41.1%, Margot Robbie with 40.9%, and Aleksej Navalnyj with 36.2%. The videos with the least "uncertain" responses were Melania Trump with 13.9%, Kim Jung Un with 11.3%, and Queen Elizabeth II with 7.8%.

There were precisely two respondents who had 100% correct answers. The average number of correct answers was 9, leading to an average success rate of 57.3%. Success rate refers to the rate of which participants have given correct answers. 57.3% is slightly higher than randomly guessing but significantly poorer than deepfake detection by today's state-of-the-art detection software (Schetinger et al. 2017, p. 142; Skibba 2020, p. 1339). When classifying real videos, participants had an average success rate of 66.04%, whereas the success rate when classifying deepfakes was only 48.59%. The average number of incorrect responses was 3, and the average uncertain responses were 4. The mean rate of uncertainty for all videos was 24.5%, although some videos have rates quite a lot higher than that. Two videos had a rate of uncertainty of 41.1% and 40.9%. These rates are significantly higher than the ones of Khodabakhsh et al. (2019), where the rates of uncertainty never exceeded 25%.

The median of the time spent completing the survey is 18 minutes and 1 second, the shortest time spent was 9 minutes and 55 seconds, and the longest response time was 1 hour, 1 minute, and 13 seconds.

5.2 Linear regression analysis

A regression analysis was considered the best option to analyze some of the results from the survey. A linear regression models the relationship between two or more variables by finding the line of best fit (Lewis-Beck, 1993). This will then allow us to predict the value of the dependent variable by the value of the independent variable. Using a regression analysis allows us to describe the strength and direction of a correlation and quantify this (Skog, 2017; Flatt and Jacobs, 2019). Finally, using a regression analysis allows us to inspect whether a correlation is curvilinear or not and whether there is statistical interaction between several independent variables (Skog, 2017; Flatt and Jacobs, 2019).

Below, I present two models with two different dependent variables. One using Classification of deepfakes, and the second using Uncertainty when classifying. Both models are included to give more insight into which factors might affect participants' ability to correctly identify a deepfake and inspect participants' certainty when deciding. The variables gender, age, education, digital literacy, internet use per day, trust in news outlets, interest in politics, and expected performance are included in both adjusted models as control variables.

Linear regression

Classification of deepfakes	**Coef.**	**St.Err.**	**t-value**	**p-value**	**[95% Conf**	**Interval]**	**Sig**
Age	-.201	.034	-5.96	0	-.267	-.135	***
Education	-.045	.062	-0.73	.468	-.168	.077	
Digital literacy	.335	.129	2.60	.01	.082	.588	***
Trust in news	.119	.118	1.01	.313	-.113	.351	
Constant	8.951	.725	12.34	0	7.527	10.376	***
Mean dependent var		9.170	SD dependent var			2.502	
R-squared		0.081	Number of obs			682.000	
F-test		14.880	Prob > F			0.000	
Akaike crit. (AIC)		3137.745	Bayesian crit. (BIC)			3160.370	

*** *p<.01, ** p<.05, * p<.1*

Model 1.1: Primary linear regression model with Classification of deepfakes as the dependent variable.

From model 1.1, we see that the mean number of correct answers was 9.170, with a standard deviation of approximately 2.5. The model has an R^2-value of 0.081, meaning that the independent variables can explain approximately 8.1% of the change in Classification of deepfakes.

Further, age and digital literacy are statistically significant. Since age has a negative sign, age decreases as the number of correct answers increases, or more simply, younger participants have a higher number of correct responses than older participants. Next, digital literacy has a positive sign, meaning there is a positive correlation between expected performance and the number of correct answers. Participants who responded with an increased level of digital literacy also have an increased number of correct answers. Both variables are statistically significant at the 1% level.

Linear regression

Uncertainty when classifying	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
Age	.184	.038	4.90	0	.11	.258	***
Education	.132	.07	1.89	.06	-.005	.269	*
Digital Literacy	-.331	.144	-2.30	.022	-.614	-.048	**
Trust in news	-.109	.132	-0.82	.411	-.367	.15	
Constant	3.636	.811	4.49	0	2.044	5.227	***

Mean dependent var	3.767	SD dependent var	2.765
R-squared	0.061	Number of obs	682.000
F-test	10.982	Prob > F	0.000
Akaike crit. (AIC)	3289.001	Bayesian crit. (BIC)	3311.626

*** p<.01, ** p<.05, * p<.1*

Model 2.1: Primary linear regression model with the number of uncertain answers as the dependent variable.

From model 2.1, we can see that the average number of uncertain answers is 3.77, with a standard deviation of approximately 2.77. The independent variables explain approximately 6.1% of the change in uncertain answers. In model 2.1, we see that in this model as well, Age and Digital literacy are statistically significant. However, the correlations have changed direction. Here, the correlation with Age is now positive, meaning that the number of uncertain answers increases with age. Digital literacy, however, is negative, meaning that participants with decreased levels of digital literacy had an increased number of uncertain responses. Age being significant at the 1% level, whereas Digital literacy is statistically significant to the 5% level.

5.3 Improved regression models

In models 1.1 and 2.1, only two variables showed a statistically significant effect on the dependent variables. However, by inspecting each variable and making a regression line, it was clear that other variables needed to be included, and some needed to be re-specified to find the best-suited regression model (see attachments 3 and 4 for a visual representation of the regression lines). The control variables Male, Expected performance, Political interest, and Internet use were added, and the variables Age, Internet use, and Education were specified differently. I will argue that the adjusted models 1.2 and 2.2 are considered the best representations of the results from the data set.

Linear regression

Classifying of deepfakes	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
Man	.236	.2	1.18	.24	-.157	.629	
Education	-.044	.065	-0.67	.504	-.172	.084	
Digital literacy	.171	.145	1.17	.241	-.115	.456	
Internet use	.999	.44	2.27	.023	.135	1.862	**
Trust in news	.088	.12	0.73	.467	-.149	.324	
Interest in politics	-.005	.067	-0.08	.939	-.137	.127	
Expected performance	.357	.134	2.67	.008	.095	.619	***
Internet use^2	-.173	.074	-2.35	.019	-.317	-.029	**
Age, ref. group: 26-30 years old							
18-22	.198	.664	0.30	.766	-1.105	1.501	
23-25	.06	.435	0.14	.89	-.794	.914	
31-35	-.452	.458	-0.99	.324	-1.351	.448	
36-40	-.416	.435	-0.96	.339	-1.271	.439	
41-45	-.962	.427	-2.25	.025	-1.8	-.124	**
46-50	-1.094	.413	-2.65	.008	-1.904	-.283	***
51-55	-1.339	.419	-3.20	.001	-2.163	-.516	***
56-60	-1.419	.485	-2.93	.004	-2.372	-.467	***
61-65	-1.404	.471	-2.98	.003	-2.33	-.479	***
66-70	-1.385	.553	-2.51	.012	-2.471	-.3	**
>70	-2.2	.607	-3.62	0	-3.392	-1.008	***
Constant	6.782	.93	7.29	0	4.955	8.609	***

Mean dependent var	9.176	SD dependent var	2.498
R-squared	0.107	Number of obs	676.000
F-test	4.149	Prob > F	0.000
Akaike crit. (AIC)	3118.448	Bayesian crit. (BIC)	3208.772

*** *p<.01, ** p<.05, * p<.1*

Model 1.2: Adjusted linear regression model with the number of correct answers as the dependent variable.

In model 1.2, R^2 has increased from 9.5% in model 1 to 10.7%. Ergo, the independent variables now explain 10.7% of the variety in the number of correct answers. Expected performance is still statistically significant with a positive sign, meaning that respondents who reported high values on expected performance also have high numbers of correct answers, significant at the 1% level.

Moreover, we see that internet use per day is statistically significant and that there is added an internet use^2 variable to explain the curvilinearity of the correlation. Because the coefficient of internet use has a positive sign, while the coefficient of internet use^2 has a negative sign, the curve initially increases but then flattens and might decrease. Further, age is represented in model 1.2 as a dummy-set, with 26-30 years old is set as a reference group. A dummy-set is different from other independent variables. Instead of looking at the influence of the independent variable as a whole, the variable is divided by each category, allowing for comparison of influence between the categories (Ringdal, 2013). Looking at the category's P-values, we see that the age groups 41-45, 46-50, 51-55, 56-60, 61-65, 66-70, and >70 are statistically significant with negative signs. Thus, these age groups have a lower number of correct answers than the reference group 26-30 years old. Hence, there is a statistically significant difference between these age groups and the reference groups, showing that 26–30-year-olds had an increased number of correct answers. Most of the mentioned age groups were statistically significant at the 1% level, whereas 41-45 and 66-70-year-olds were statistically significant to the 5% level.

Linear regression

Uncertainty when classifying	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
Gender (Man)	-.278	.223	-1.25	.212	-.716	.159	
Digital literacy	-.056	.163	-0.34	.732	-.377	.265	
Internet use	-1.379	.491	-2.81	.005	-2.343	-.414	***
Trust in news	-.058	.134	-0.43	.665	-.321	.205	
Interest in politics	.098	.075	1.31	.19	-.049	.245	
Expected performance	-.395	.149	-2.65	.008	-.687	-.103	***
Internet use^2	.222	.083	2.69	.007	.06	.384	***
Education, ref. group: higher education, no degree							
Primary and secondary school	1.445	1.092	1.32	.187	-.701	3.59	
High school	.94	.519	1.81	.071	-.08	1.959	*
Higher education, one year	.551	.559	0.99	.325	-.547	1.65	
Higher education, bachelor's degree	1.185	.477	2.48	.013	.248	2.121	**
Higher education, master's degree	1.139	.484	2.35	.019	.189	2.09	**
Higher education, Ph.D.	1.476	.627	2.35	.019	.245	2.707	**
Age, ref. group: 51-55 years old							
18-22	-1.953	.722	-2.71	.007	-3.37	-.536	***
23-25	-1.17	.451	-2.59	.01	-2.056	-.284	***
26-30	-1.42	.47	-3.02	.003	-2.343	-.497	***
31-35	-1.306	.467	-2.80	.005	-2.223	-.389	***
36-40	-1.076	.435	-2.47	.014	-1.93	-.221	**
41-45	-.773	.421	-1.84	.067	-1.599	.054	*
56-50	-.702	.405	-1.73	.084	-1.497	.094	*
56-60	-.452	.484	-0.93	.351	-1.403	.499	
61-65	.322	.454	0.71	.479	-.57	1.213	
66-70	.077	.557	0.14	.89	-1.016	1.17	
>70	-.519	.62	-0.84	.403	-1.736	.699	
Constant	6.649	1.115	5.96	0	4.459	8.839	***

Mean dependent var	3.751	SD dependent var	2.752
R-squared	0.107	Number of obs	676.000
F-test	3.247	Prob > F	0.000
Akaike crit. (AIC)	3259.657	Bayesian crit. (BIC)	3372.561

*** *p<.01, ** p<.05, * p<.1*

Model 2.2: Adjusted linear regression model with the number of uncertain answers as the dependent variable.

Model 2.2 shows that R^2 has increased from 7.5% in model 2.2 to 10.7%. Thus, the independent and control variables explain approximately 11% of the change in Uncertainty when classifying. Further, Internet use is statistically significant with a negative sign, and Internet use^2 has been added, being statistically significant with a positive sign. The squared variable was added to explain the curvilinearity further, and by the signs of the coefficients, the curve initially drops before flattening and might eventually increase. Expected performance is also statistically significant to the 1% level with a negative sign, meaning that there is a negative correlation between expected performance and the number of uncertain answers. Therefore, participants who reported low expected performance had an increased number of uncertain answers.

Moreover, the education variable comes in the shape of a dummy-set with higher education - no degree being the reference group. From the dummy-set, we see that participants with

bachelor's degrees, master's degrees, or Ph.D.'s have an increased number of uncertain answers than participants without a degree, which is statistically significant to the 5% level. A dummy set of the age variable is also included to show a more detailed representation of the differences between age groups. In this model, the reference group was set to 51-55 years old, and it shows that the age groups 18-40, in general, has a lower number of uncertain answers than the 51-55-year-olds.

When evaluating the results of regression analysis, certain assumptions should be fulfilled to assess the robustness of the results. In the next sub-chapter, I will detail what these assumptions are, whether the data met the required assumptions, and the potential consequences of not meeting them.

5.4 Assumptions for linear regression analysis

The assumptions of linear regression include that the residual should be statistically independent, and their variation should be homoscedastic and normally distributed. Additionally, the residuals should not be autocorrelated and there should be an absence of multicollinearity (Skog, 2017; Mehmetoglu and Jakobsen, 2017; Flatt and Jacobs, 2019). Not meeting these assumptions might lead to biased and misleading projections if the violations are substantial. However, minor violations might be tolerable depending on the model's applications (Flatt and Jacobs, 2019).

The first assumption states that the residuals should not be autocorrelated, meaning that the residuals should be uncorrelated for different observations (Mehmetoglu and Jakobsen, 2017). This, however, should not be a problem when using a random sample from a population and when data is only collected once, not periodically.

The following assumption is for the residuals to be homoscedastic, which means that the variation around the regression line is relatively equal for all values of the independent variable (Skog, 2017). Not meeting this assumption means that the residuals are heteroscedastic, and that the model's predictability varies with the values of the variable. Heteroscedasticity does not affect R^2 or adjusted R^2 but will, affect standard deviation, t-values, f-values, and confidence intervals. However, if the number of observations is high, a small amount of heteroscedasticity will not be a problem (Ringdal, 2013). Regression model 1.2 with Classification of deepfakes shows homoscedastic residuals, whereas model 2.2 does not.

Correspondingly, the residuals should also be normally distributed. If this assumption is not met, the P>t values will not be valid. This might affect the ability to determine the significance of model coefficients (Flatt and Jacobs, 2019). Attachment 5 illustrates the variable’s histogram and reveals relatively normally distributed residuals. Testing shows that the variables Education, Internet use, Trust in news, and Interest in politics do not have normally distributed residuals, whereas Man and Digital literacy do. Despite the assumption that the residuals should be normally distributed, research from real-life situations rarely consists of perfectly normally distributed residuals, and a close proximate should be sufficient (Flatt and Jacobs, 2019).

A local polynomial regression can indicate which variables do not meet the assumption of linearity (see attachments 3 and 4 for local polynomial regressions for each independent variable) (Mehmetoglu and Jakobsen, 2017). From both attachments 3 and 4, it is evident that variables like Internet use per day, Age, and Education are variables that do not fit a linear correlation with either the number of correct answers or the number of uncertain answers. This problem can be solved by squaring the variables, making them curvilinear,

or making a dummy set where one category is set as a reference (Skog, 2017). In models 1.2 and 2.2, I have taken the variables that showed signs of curved linearity in models 1.1 and 2.1 and adjusted them to give a more accurate description of their effect on the dependent variables.

The Pearson correlation using Classification of deepfakes shows that no variables show a strong correlation, with Internet use and Digital literacy having the strongest correlation value of 0.358. Model 2.2 with Uncertainty when classifying as dependent variable showed similar results, with Internet use and Digital literacy and Internet use and Age having the strongest correlations with a Pearson's R of 0.358 and -0.333, respectively.

Another assumption for linear regression is that the independent variables and the residuals are uncorrelated. Hence there should be no underlying causal factors for the dependent variable that correlates with the independent variable (Skog, 2017). If this assumption is not met, the model will give a wrong impression of how our independent variables affects the dependent variable. This is called omitted variable bias, and indications of this include unexpected signs on variable coefficients, non-significant constant, and that adjusted R^2 increases considerably when a presumed omitted variable is included in the model. However, this only affects the model if another variable correlates with both the dependent- and the independent variable (Skog, 2017). Adding relevant independent variables to the model decreases the risk of not meeting this assumption.

Lastly, there should be an absence of multicollinearity, meaning that two independent variables perfectly correlate. If this occurs, your model has two variables that, in essence, measure the same phenomenon, which will lead to standard errors being too small (Mehmetoglu and Jakobsen, 2017). Testing showed that Model 2.2 with Classification of deepfakes showed no multicollinearity. In contrast, model 2.2, using Uncertainty when classifying, showed that a category in the Education dummy-set shows signs of multicollinearity. This means that this category, Higher education with a bachelor's degree, measures approximately the same as the reference category, being Higher education without a degree. Although this result shows borderline collinearity, it is reasonable that these categories show somewhat similar results. However, because the values are just within the desired value boundaries, the categories are kept as is instead of making an index.

To summarize, the residuals seem to be uncorrelated. Certain variables show a linear correlation of varying extents, whereas others show a curvilinear correlation. However, heteroscedasticity is present in the residuals of the model using Uncertainty when classifying as the dependent variable, whereas not in the model using the number of correct answers. Models 1.2 and 2.2 are correctly specified, but they lack explanatory variables. Finally, there is no significant multicollinearity in either model. One solution to heteroscedasticity and non-normality could be to logarithmically transformation variables. Transforming variables that contained heteroscedasticity improved them, but the variables still did not show statistic significant influence. Hence, variables remained unchanged in the analysis.

5.5 Logistic regression: previous knowledge and previously seen video

In this sub-chapter, I will analyze the influence of the variables Previous knowledge of the video and Previous knowledge of the subject on correctly classifying a video and giving an uncertain answer. For this analysis, logistic regression was considered best suited. In

logistic regression, the dependent variable is a dummy variable containing only two categories, 0 and 1. When using logistic regression, we calculate the probability of the dependent variable having the value 1 (Mehmetoglu and Jakobsen, 2017; Stoltzfus, 2011). Model 3 represents the logistic regressions for each video with the odds ratio for each group of questions.

In model 3, each name (in bold) represents the variable concerning the correct classification of the video in question. These variables are different from Classification of deepfakes as used in the linear regression, as these only concern each video. In contrast, Classification of deepfakes was a variable generated by adding all classifying variables together. The different subjects/videos have individual logistic regression models as they are independent of each other, meaning that, for instance, having previous knowledge of Stefan Löfven does not affect your ability to identify a video of Zendaya correctly.

In addition to coefficients, the model also includes estimated odds ratios because this allows for a more detailed calculation of the odds of being in group y=1 when x goes from 0 to 1. Alternatively, it allows us to calculate the odds of a respondent giving a correct answer when they have either previously seen the video or have previous knowledge of the subject. Using the formula (OR-1)*100, we get the percentage change in odds. The percentage change in odds for each variable can be seen in the second column of the model.

Logistic regression – correct answers

	Coef.	**(OR-1)*100**	**St.Err.**	**t-value**	**p-value**
Angela Merkel					
Previous knowledge of video	.788	119.9	0.84	0.94	.349
Previous knowledge of subject	-.095	-9	.345	-0.27	.784
Constant	.223		.419	0.67	.506
Pseudo r-squared	0.001				
Prob>chi2	0.599				
Boris Johnson					
Previous knowledge of video	.147	15.9	.507	0.29	.771
Previous knowledge of subject	.423	52.6	.432	0.98	.328
Constant	.865		.421	2.05	.04
Pseudo r-squared	0.001				
Prob>chi2	0.602				
Aleksej Navalnyj					
Previous knowledge of video	-.660	-48.3	.918	-0.72	.472
Previous knowledge of subject	.240	27.1	.157	1.53	.127
Constant	.015		.121	0.12	.904
Pseudo r-squared	0.003				
Prob>chi2	0.257				
River Phoenix					
Previous knowledge of video	-.188	-17.1	.312	-0.60	.547
Previous knowledge of subject	.219	24.5	.192	1.14	.253
Constant	-.987		.106	-9.33	0
Pseudo r-squared	0.002				
Prob>chi2	0.513				
Queen Elizabeth					
Previous knowledge	.217	24.2	.540	0.40	.688

of video					
Previous knowledge of subject	.176	19.2	.633	0.28	.782
Constant	1.837		.622	2.95	.003
Pseudo r-squared	0.000				
Prob>chi2	0.884				
Zendaya					
Previous knowledge of video	-1.745	-82.5	1.264	-1.38	.168
Previous knowledge of subject	1.523	358.8	.328	4.64	0
Constant	.915		.095	9.63	0
Pseudo r-squared	0.040				
Prob>chi2	0.000				
Stefan Löfven					
Previous knowledge of video	1.345	283.9	.423	3.18	.001
Previous knowledge of subject	.334	39.7	.190	1.76	.079
Constant	-.043		.167	-0.25	.802
Pseudo r-squared	0.019				
Prob>chi2	0.000				
Steve Carell					
Previous knowledge of video	1.617	403.6	.348	4.64	0
Previous knowledge of subject	1.257	251.6	.206	6.12	0
Constant	-2.036		.146	-13.94	0
Pseudo r-squared	0.111				
Prob>chi2	0.000				
Dua Lipa					
Previous knowledge of video	1.434	319.4	1.128	1.27	.204
Previous knowledge of subject	.499	64.7	.178	2.80	.005
Constant	-.435		.091	-4.78	0
Pseudo r-squared	0.012				
Prob>chi2	0.005				
Margot Robbie					
Previous knowledge of video	1.846	533.6	.477	3.87	0
Previous knowledge of subject	.952	159.2	.179	5.32	0
Constant	-1.103		.106	-10.39	0
Pseudo r-squared	0.064				
Prob>chi2	0.000				
Jared Leto					
Previous knowledge of video	1.703	449.2	1.056	1.61	.107
Previous knowledge of subject	.001	0.1	.162	0.01	.993
Constant	.598		.108	5.52	0
Pseudo r-squared	0.005				
Prob>chi2	0.130				
Edward Snowden					
Previous knowledge of video	1.776	490.4	.524	3.39	.001
Previous knowledge of subject	.252	28.7	.240	1.05	.294
Constant	.956		.215	4.45	0
Pseudo r-squared	0.030				
Prob>chi2	0.000				
Kim Jung-Un					

Previous knowledge of video (omitted)	1	0	.	.	.
Previous knowledge of subject	.789	120.1	.368	2.15	.032
Constant	1.099		.348	3.16	.002
Pseudo r-squared	0.008				
Prob>chi2	0.043				
Melania Trump					
Previous knowledge of video	.835	130.5	.624	1.34	.181
Previous knowledge of subject	1.080	194.5	.274	3.95	0
Constant	.182		.256	0.71	.477
Pseudo r-squared	0.023				
Prob>chi2	0.000				
Nicolas Cage					
Previous knowledge of video	1.627	409.1	.433	3.76	0
Previous knowledge of subject	.690	99.4	.264	2.62	.009
Constant	-.920		.251	-3.67	0
Pseudo r-squared	0.028				
Prob>chi2	0.000				
Mette Frederiksen					
Previous knowledge of video	-1.513	-78	.542	-2.79	.005
Previous knowledge of subject	.207	23	.172	1.20	.229
Constant	-.664		.141	-4.70	0
Pseudo r-squared	0.013				
Prob>chi2	0.003				

Model 3: Logistic regression with correct identification as the dependent variable.

From model 3, we might first and foremost notice that the models referring to the videos of Angela Merkel, Boris Johnson, Aleksej Navalnyj, River Phoenix, Queen Elizabeth, and Jared Leto contain no significant values and can be disregarded. Next, the models show a noticeably high percentage change in odds for some variables, with values like 533%, 490%, and 409%. Such high values are because all variables in the model only have two categories, 0 and 1. Hence, all 682 participants are located either in 0 or in 1. If an increased number of participants have given the same response to specific questions, it will result in significantly high values of percentage change in odds when the independent variable goes from 0 to 1.

Moreover, we can see that in the video portraying Zendaya, participants who had previous knowledge about her had 358.8% higher odds for correctly identifying the video as genuine than the ones with no previous knowledge. This also being statistically significant to the 1% level. In the video of Stefan Löfven, we see that participants who had seen the video or other versions of the video had 283.9% higher odds of correctly identifying the video as genuine, being statistically significant to the 1% level.

Further on, we see that in the video portraying Steve Carell, respondents who had previously seen the video had 403.6% higher odds to correctly identify the video as fake, and participants with previous knowledge of Steve Carell had 251.6% higher odds of doing so. Both being significant to the 1% level. In Dua Lipas video, participants who had previous knowledge of her had 64.7% higher odds of correctly identifying the video as fake than those without previous knowledge. For Margot Robbie, participants who had

previously seen the video had 533.6% higher odds of correctly identifying the video as fake, whereas participants with previous knowledge of her had 159.2% higher odds to do so. Both variables being significant to the 1% level.

The video of Edward Snowden showed that participants who had previously seen the video had 490.4% higher odds of correctly identifying the video as genuine. Further on, we see that the model of Kim Jung-Un is missing a variable. This is because the variable Previous knowledge of video showed a perfect correlation with the dependent variable. This means that all of the participants who had previous knowledge of the video of Kim Jung-Un managed to correctly identify the video as fake.

When looking at Melania Trump, we see that having previous knowledge of her increased the odds of correctly identifying the video as genuine with 194.5%. Similarly, having previously seen the video of Nicolas Cage increased the odds of correctly identifying it as genuine by 409.1%, whereas having previous knowledge of him increased the odds by 99.4%. Lastly, having previous knowledge of the video of Mette Frederiksen decreased the odds of correctly identifying the video as fake by -78%.

Although previous knowledge has various effects on participants' abilities to detect deepfakes, previous research has shown that it may affect participants' confidence when answering (Khodabakhsh et al., 2019). Research has shown that having previous knowledge of the subject in the video might decrease the number of uncertain responses. Model 4 portrays a logistic regression model with uncertain responses as dependent variables for all 16 videos, including odds ratios and percentage change in odds.

Logistic regression – uncertain answers

	Coef.	(OR-1)*100	St.Err.	t-value	p-value
Angela Merkel					
Previous knowledge of video	-.456	-36.6	1.085	-0.42	.674
Previous knowledge of subject	.743	110.3	.539	1.38	.168
Constant	-2.079		.530	-3.92	0
Pseudo r-squared	0.003				
Prob>chi2	0.296				
Boris Johnson					
Previous knowledge of video	-1.492	-77.5	1.026	-1.45	.146
Previous knowledge of subject	-.474	-37.7	.476	-1.00	.32
Constant	-1.253		.463	-2.71	.007
Pseudo r-squared	0.008				
Prob>chi2	0.109				
Aleksej Navalnyj					
Previous knowledge of video	-.767	-53.5	1.123	-0.68	.495
Previous knowledge of subject	-.144	-13.4	.162	-0.88	.376
Constant	-.476		.124	-3.83	0
Pseudo r-squared	0.002				
Prob>chi2	0.494				
River Phoenix					
Previous knowledge of video	-1.110	-67	.426	-2.60	.009
Previous knowledge of subject	-.719	-51.3	.199	-3.61	0

Constant	-.413		.096	-4.29	0
Pseudo r-squared	0.030				
Prob>chi2	0.000				
Queen Elizabeth II					
Previous knowledge of video	-.073	-7	.618	-0.12	.906
Previous knowledge of subject	.587	79.8	1.034	0.57	.57
Constant	-3.041		1.024	-2.97	.003
Pseudo r-squared	0.001				
Prob>chi2	0.820				
Zendaya					
Previous knowledge of video (omitted)	1		.	.	.
Previous knowledge of subject	-2.420	-1.21	.594	-4.08	0
Constant	-1.375		.107	-12.87	0
Pseudo r-squared	0.058				
Prob>chi2	0.000				
Stefan Löfven					
Previous knowledge of video	-1.379	-74.6	.608	-2.25	.024
Previous knowledge of subject	-.094	-9	.222	-0.42	.672
Constant	-1.127		.195	-5.79	0
Pseudo r-squared	0.011				
Prob>chi2	0.020				
Steve Carell					
Previous knowledge of video	-2.424	-91.1	.734	-3.30	.001
Previous knowledge of subject	-.838	-56.8	.181	-4.63	0
Constant	-.016		.095	-0.17	.864
Pseudo r-squared	0.059				
Prob>chi2	0.000				
Dua Lipa					
Previous knowledge of video (omitted)	1		.	.	.
Previous knowledge of subject	-.094	-9	.196	-0.48	.632
Constant	-.839		.097	-8.67	0
Pseudo r-squared	0.000				
Prob>chi2	0.630				
Margot Robbie					
Previous knowledge of video	-2.086	-87.6	.741	-2.81	.005
Previous knowledge of subject	-.581	-44.1	.181	-3.20	.001
Constant	-.148		.092	-1.61	.108
Pseudo r-squared	0.031				
Prob>chi2	0.000				
Jared Leto					
Previous knowledge of video (omitted)	1		.	.	.
Previous knowledge of subject	-.165	-15.2	.190	-0.87	.384
Constant	-1.217		.123	-9.85	0
Pseudo r-squared	0.001				
Prob>chi2	0.383				
Edward Snowden					
Previous knowledge of video	-1.924	-85.4	.727	-2.65	.008

Previous knowledge of subject	-.525	-40.8	.256	-1.98	.047
Constant	-1.253		.231	-5.41	0
Pseudo r-squared	0.034				
Prob>chi2	0.000				
Kim Jung-Un					
Previous knowledge of video (omitted)	1		.	.	.
Previous knowledge of subject	-.765	-53	.395	-1.91	.056
Constant	-1.358		.374	-3.63	0
Pseudo r-squared	0.007				
Prob>chi2	0.073				
Melania Trump					
Previous knowledge of video (omitted)	1		.	.	.
Previous knowledge of subject	-1.219	-70.4	.299	-4.07	0
Constant	-.718		.273	-2.63	.008
Pseudo r-squared	0.027				
Prob>chi2	0.000				
Nicolas Cage					
Previous knowledge of video	-2.779	-93.8	1.019	-2.73	.006
Previous knowledge of subject	-.484	-38.4	.248	-1.95	.051
Constant	-.270		.231	-1.17	.243
Pseudo r-squared	0.027				
Prob>chi2	0.000				
Mette Frederiksen					
Previous knowledge of video	-.969	-62	.616	-1.57	.116
Previous knowledge of subject	-.329	-28	.193	-1.70	.088
Constant	-1.028		.152	-6.78	0
Pseudo r-squared	0.010				
Prob>chi2	0.033				

Model 4: Logistic regression with uncertain classification as the dependent variable.

From model 4, we can see that previous knowledge of the video or the subject in the video does not have a significant effect when it comes to the video portraying Angela Merkel, Boris Johnson, Aleksej Navalnyj, Queen Elizabeth II, or Mette Frederiksen.

For the videos portraying River Phoenix, Steve Carell, Margot Robbie, and Edward Snowden, both having previous knowledge of the video and the subject in the video reduced the odds of giving an uncertain answer by 40.8-91.1 percent. In the videos portraying Zendaya, Dua Lipa, Jared Leto, Kim Jung-Un, and Melania Trump, the variable concerning previous knowledge of the video is omitted in the model because of perfect collinearity. Ergo, no participants who responded that they had previous knowledge of the video gave an uncertain answer. For Zendaya and Melania Trump, having previous knowledge of the subject significantly affected and decreased the odds of giving an uncertain response by 1.21 and 70.4 percent, respectively. Previous knowledge of the subject did not affect the videos of Dua Lipa, Jared Leto, and Kim Jung-Un.

For the videos portraying Stefan Löfven and Nicolas Cage, previous knowledge of the video was statistically significant, reducing the odds of giving an uncertain response by 74.6 and 93.8 percent, respectively, whereas previous knowledge of the subject was not. Like linear

regression, logistic regression also comes with some assumptions that should be met to ensure the quality of the models. In the following sub-chapter, I will explain some of these assumptions, whether my dataset meets them, and the potential consequences of not meeting them.

5.6 Assumptions of logistic regression

The assumptions for logistic regression include that the model must be correctly specified, no irrelevant variables should be in the model, the observations need to be independent of each other, and the absence of multicollinearity (Mehmetoglu and Jakobsen, 2017).

The assumption that the variables should be correctly specified is mainly based on theory. However, this can also be tested using a Hosmer-Lemeshow goodness-of-fit test (Mehmetoglu and Jakobsen 2017, p. 70). Performing a goodness-of-fit test showed that all models, except Kim Jung-Un and the variable Previous knowledge of video, were correctly specified.

Moreover, there should be no irrelevant variables in the models. Whether a variable is considered irrelevant is based on both theoretical background and statistical significance (Mehmetoglu and Jakobsen, 2017). Considering the variables were included based on previous research, I will argue that the models contain no irrelevant variables. The assumption that the observations should be independent of each other was also discussed in the assumptions for linear regression. Because the data was collected from a random sample and was only collected once, the observations can be considered independent of each other (Skog, 2017).

Lastly, there should be no multicollinearity. However, this test cannot be run on logistic regressions, but using linear regression instead of logistic regression allows to test for multicollinearity. All models passed the test, meaning no multicollinearity between the independent variables.

5.7 Factors of distinction

The last variable that will contribute to a broader understanding of how Norwegians evaluate audiovisual content online is by looking at the cues used to decide when classifying videos. Figure 4 shows the distribution of the cues from least to most used on the real videos and the deepfakes.

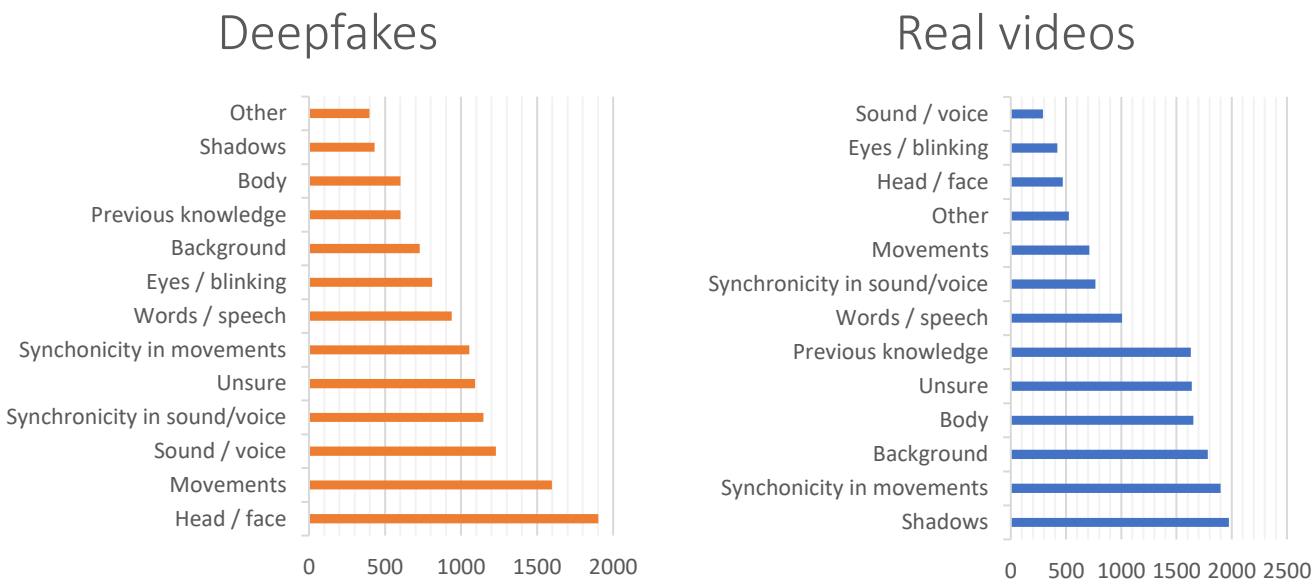


Figure 4: Factors affecting participant's decisions.

What is interesting to see in Figure 4 is that the "unsure" category is highly used in both real videos and in deepfakes, ranking at the number 5. Further, we see that the distribution of clues differed depending on whether the video was genuine or fake. For deepfakes, the most used cues were head/face, movements, sound/voice, and synchronicity in sound/voice. Moreover, the most used cues in real videos were shadows, synchronicity in movements, background, and body.

In addition to the factors seen in Figure 4, respondents were able to choose the "other"-option which allowed them to write another answer in a text box. Each video had quite a few of such responses that have not been included in the Figure. Some of the most common additional responses were that there were glitches in the video, uncertainty around whether the subject in the video would say what they did, hair, lips, that they lack a sender in order to evaluate the credibility of the video, and that they have no knowledge of the subject to make a decision.

Like written above, the two most obvious videos that had the highest share of correct answers were the one portraying Queen Elizabeth II and Kim Jong-Un. For Queen Elizabeth II, the cues mentioned the most in the textboxes were the colors of the video, her unnatural voice, and a general creepiness of the video. For Kim Jong-Un, the most used cues besides the ones given, were the disbelief that he could speak English and the content of what he said. The two videos that were hardest to classify was Steve Carell and River Phoenix. For Steve Carell, the most used written cues were that there was no speech to evaluate from, that some had seen the film and knew that someone else was the real actor, and that they had no previous knowledge of the movie. For River Phoenix, the most used written cues were that they had previously seen the episode and knew who the original actor was, lips, and asynchrony in the voice.

6 Discussion

The main scope of the thesis is to examine whether a Norwegian sample can detect deepfakes, which factors affect this ability, and whether they do so with confidence or with uncertainty. The discussion will first examine the question regarding Norwegian's ability to detect deepfakes before examining the hypotheses in the order they were mentioned above. Lastly, I will give a summarizing discussion with some reflections and limitations.

6.1 To what extent are Norwegians able to recognize deepfakes?

The overall research question asked to which extent Norwegians can identify deepfakes, and what factors influence this. The respondents had an overall success rate of 57.3% or nine out of 16 correct answers. Participants correctly identified real videos with a success rate of 66.04% and deepfake videos with a success rate of 48.59%. A success rate of 57.3% seems to coincide adequately with previous research, as the deepfakes included in the survey were made using the most sophisticated manipulation tools (Korshunov and Marcel, 2020; Vaccari and Chadwick, 2020; Khodabakhsh et al., 2019). This performance, however, can be considered dissatisfactory compared to deepfake detection software, with success rates from 65-80% (Skibba, 2020). However, a remarkable finding in the research is that only two out of 682 respondents had a success rate of 100% when classifying the videos. Considering that Norwegians, in general, are considered to have high levels of digital literacy compared to other countries, I expected this number to be higher.

This result implies that the videos were hard to distinguish, or that participants were unaware of which cues could be useful to detect a deepfake when they completed the survey. This implies that the overall awareness of what deepfakes is among the Norwegian population is low.

In this project, both the deepfakes and the real videos were kept at a 480p resolution. This might have influenced the participants' ability to classify correctly since all videos were quite compressed. Deepfakes are often highly compressed videos because the technology is not yet able to make perfectly synthesized videos of high quality, and one way to increase the probability of deceiving the audience is to compress the video to make the flaws less obvious. Another possible factor why participants were struggling with correctly classifying the videos is that essential information that would have been provided in a natural situation was missing. Information such as the sender and content with a clear message are essential cues when critically evaluating the authenticity of the information. When this is not provided, it seems that Norwegians struggle with deciding the videos' authenticity.

Admittedly, the survey did not manage to include respondents with low digital literacy. The level of digital literacy in the sample is high, with the mean response being "quite high digital literacy." However, considering this low number of perfect scores, perhaps Norwegians are not as digitally literate as first assumed. Moreover, although digital literacy includes the ability to evaluate and process digital content, perhaps deepfakes are not considered as such digital content. It could be that the number of needed skills needs to be broadened (DigComp 2.0, 2019).

Further, we see in the analysis that there are significant differences in which videos were classified correctly with ease and which were more complex. The video portraying Queen

Elizabeth II of England was the video with the highest number of correct responses. 602 participants correctly classified it as a deepfake. Several factors might cause this. Firstly, Queen Elizabeth II is a world-famous celebrity, and as previously noted, knowing the subject decreases uncertainty when responding. 660 participants responded that they had previous knowledge of the Queen, which might have influenced the results substantially. Further, the deepfake portraying the Queen was initially made by VFX studio Framestore for Channel 4 as an alternative Christmas message and a parody of the Queen's previous Christmas Broadcasts by BBC, and it went viral (Rahim, 2020; BBC, 2019; Channel 4, 2020).

Because of this, we might expect that several respondents had seen the video before participating in the survey. However, only 41 respondents replied that they had seen the video before, and 35 replied that they were uncertain. Hence, this implies that the design of the survey might be causing these results. Moreover, the deepfake of Queen Elizabeth II might be considered the video with the most apparent look of a deepfake. Because the video was intended to be a parody, the goal was probably not to make the most realistic video. In the deepfake, the room is almost unnaturally bright, with pictures of several royal family members on the desk. Although in BBC's real video of the Queen's speech, you can see a photo of Prince Phillip, the deepfake also includes photos of their dog, Meghan Markle and Prince Harry, Prince Charles and Camilla, among others (BBC, 2019). Photos like these would most likely not have been included in an actual Christmas speech.

Another video that had a high number of correct responses was the video portraying Kim Jong-Un. 588 respondents correctly classified the video as a deepfake. In the deepfake, Kim Jong-Un is speaking somewhat broken English. On the question asking about the most important cues when deciding, quite a few respondents wrote that they did not believe he could speak English. However, the real Kim Jong-Un went to school in Switzerland, where he was taught English (Murray, 2020). Hence, although they correctly classified the video, it might have been based on the wrong assumptions. This coincides with what Nightingale et al. (2017) also found in their research that respondents did manage to identify the content correctly but tend to mislocate the manipulation location. Hence, their correct classification might be based on the wrong reasons, and the results might have been different if he spoke Korean in the video.

Conversely, another reason why the video of Kim Jong-Un was one of the most correctly classified videos might be because of what he is referencing. In the deepfake, he talks about closed voting locations, leading to millions being unable to vote. Considering North Korea is a dictatorship, talking about democratic processes like voting could be considered strange. This might have contributed strongly to the high success rates when classifying this deepfake.

Moreover, another reason why Queen Elizabeth II and Kim Jong-Un had such a high success rate, might because they are 2 out of 3 videos in the survey that included subtitles. Subtitles might be a contributing factor when classifying the videos and might be efficient tools for interpretation or misinterpretation of the videos. One particular reason why Kim Jong-Un had such a success rate could be that the subtitles included in the video shows what he was "saying" in the video before it was cropped to fit the qualifications. On the starting screen of the video, the subtitles of what was said prior states "*No democracy. I don't have to do anything; you're doing it to yourselves*." This is arguably something he would not say in public, which participants might also have considered. The third video including subtitles was the deepfake of Mette Frederiksen, which means that only

deepfakes had subtitles. However, Mette Frederiksen was the 4th hardest video to classify by the sample. Hence, the importance of subtitles might be important cues in some videos, but not others.

When looking at the videos with the least number of correct responses, the video portraying Steve Carell is at the bottom with 140 correct responses. This might be because it is one out of only two videos where the subject does not speak, and only music is present. Hence, respondents cannot evaluate the video based on speech, the sound of their voice, or the content. These are all crucial clues when evaluating the authenticity of a video. Another reason why the deepfake of Steve Carrell was hard to classify might be because the video is overall quite dark, which might make any inconsistencies in the facial area hard to see. Lighting and inconsistencies in lighting have also been studied as an essential cue to detect fake media content in previous research (Nightingale et al., 2017; Mirsky and Lee, 2021; Khodabakhsh et al., 2019). The other video without speech was the video of Dua Lipa playing the saxophone, which is the fifth place of videos with the least number of correct responses.

Further, the second video that was difficult to classify was the one portraying River Phoenix. 193 participants correctly identified the video as a deepfake. One reason for this might be that River Phoenix's looks are quite similar to the actor on which his face was substituted onto, Charlie Heaton. Hence, the changes made when creating the deepfake might not have been as substantial as they would if the actors were not similar looking. Because of this, the video might have been increasingly difficult to classify, even for the 66 respondents who reported to have previously seen the video. Additionally, the lighting in the video is quite dark, which might also have increased the difficulty of classifying the video.

Additionally, another point of discussion is that both the videos that had the most correct and the least correct responses were deepfakes. Queen Elizabeth II and Kim Jong-Un are the two videos with the highest number of correct responses in ascending order. Conversely, the five videos with the least number of correct responses were all deepfakes, portraying Steve Carell, River Phoenix, Margot Robbie, Mette Frederiksen, and Dua Lipa. This shows that respondents performed quite poorly when classifying deepfakes in general but had a high success rate on the videos that were most obvious to be deepfakes. Either because they were made as parodies or included quite evident clues. Even so, although some deepfakes might be obvious for humans, research show that those are often more difficult for algorithms to detect (Korshunov and Marcel, 2020). Hence, I argue that we need the expertise of both humans and algorithms to manage the growing threat of deepfakes.

Moreover, another possible reason why there were only two participants with perfect scores is that possibly there is a knowledge gap in the Norwegian population concerning deepfakes. It seems as though only a small portion of the population knows what they are and typical cues of what to look for to distinguish them from other real media. This knowledge gap might be caused by the fact that Norwegians are exposed to fake media to a lower extent than other countries, and hence might be unaware of their existence entirely (Newman et al., 2018).

6.2 Which factors affect the ability to detect deepfakes?

To understand which factors affect and influence the ability of respondents to classify the videos correctly, I will evaluate the hypotheses mentioned in chapter 3. Further, I will discuss potential explanations and implications the results have for this research, and for society as a whole.

6.2.1 Age

H1 stated that "*Older respondents will perform worse when classifying deepfakes than younger respondents*." With the linear regression in mind, we could see that participants aging from 41 to older than 70 years old have fewer correct answers than 26-30 years old when classifying videos. Participants aged 18-25 and 31-40 show no significant variation. This might indicate that higher age reduces the number of correct responses. This implies that older responders did not recognize the fake videos as well as the younger participants. The differences between the age groups, however, are not big. 41-70-year-olds give approximately 1 less correct answer than 26-30-year-olds, and the differences between the age groups ranging from 41-70 years are quite small. What this could mean is that at the age of approximately 40, ability to detect deepfakes decreases. The negative correlation between age and the number of correct answers is in line with previous research that suggested that higher age reduces the ability to detect fake content (Khodabakhsh et al., 2019; Schetinger et al., 2017).

The difference between age groups is also apparent when giving uncertain responses, where respondent aging from 18-40 shows significantly fewer uncertain answers than 51 to 55-year-olds. Respondents aging from 41-50 also show tendencies of giving less uncertain answers than 51-55, but the difference is not as significant as the other age groups. This shows that younger respondents could be more confident when answering the survey since they rarely respond "uncertain/do not know." The difference in age groups varies, with the most significant difference being between 18–22-year-olds who generally give approximately two less uncertain answers than 51–55-year-olds. Based on the results from the linear analysis, I will argue that age affects the ability to detect deepfakes, starting at approximately 40 years old, and that we can retain H1.

The divides between age groups might be caused by the growing focus on teaching digital literacy in school for the past 25 years (Erstad, 2006). Hence, pupils in school have been taught the essential skills to become digitally literate since 1996, whereas people attending school before 1996 have not been taught this. This might have started a knowledge gap and a general gap in digital literacy between these age groups. This knowledge gap might contribute to older people not being able to separate fake news from real news (Hwang and Jeong, 2009).

Age and digital literacy are also highly correlated with the emergence and commercialization of the internet (Siapera, 2017). Older people who did not participate in the commercialization of the internet might have struggled with having the motivation and ability to learn to use the internet in productive ways, which will directly affect their digital literacy. This might be a reason why older respondents performed poorer than younger respondents.

The difference in age groups concerning the number of uncertain answers contradicts previous research where age showed a moderate negative correlation with uncertain answers (Khodabakhsh et al., 2019). This contradiction in results might be caused by the

low number of participants in the study conducted by Khodabakhsh et al. in 2019, based on only 30 participants.

Some of the same reasons might cause the positive correlation between age and the number of uncertain answers as with correct answers: digital literacy in school. Being taught in school the most important abilities of digital literacy will also increase peoples' confidence. The consequences of not having the digital literacy needed to be confident when evaluating content online might be growing knowledge gaps.

6.2.2 Higher education

H2 postulated that "*Respondents with higher education will perform better when classifying deepfakes than respondents without higher education*." Higher education is thought to be tightly connected with digital literacy, considering that an important aspect of higher education is to teach students to critically evaluate information and sources, which are also properties of being digitally literate. However, the variable showed no significant effect on correctly identifying deepfakes. This finding is in line with previous research saying that higher education does not affect the ability to detect fake media (Khodabakhsh et al., 2019; Schetinger et al., 2017). However, although the variable is not significant when classifying deepfakes, the sign of the coefficient is negative, meaning that if the variable had been significant, the correlation had been the opposite of what H2 postulated. This is an interesting finding that goes against all previous hypotheses about the influence of education (Sande, 1989).

Interestingly, education shows significant differences within the variable categories regarding the number of uncertain answers. People with a bachelor's degree, master's degree, and a Ph.D. seem to be less confident when answering the survey, giving a higher number of uncertain responses than those with higher education but with no degree. Participants who completed high school also show tendencies of being more confident in their responses than those without a degree, but less significantly than those with degrees.

However, because the variable does not significantly influence the ability to detect deepfakes correctly, H2 can be discarded.

One reason why education showed no significant effect on detecting deepfakes could be that the respondents lacked crucial information about the sender and the content of what each video said. As mentioned above, students should be taught critical evaluation based on information sources and not based on looks. Hence, if the research had been conducted in a more realistic environment and had given more information about the videos, the variable might have given more significant results.

However, the fact that education did not influence respondent's performance when detecting deepfakes is a favorable finding. Consequently, that there are no significant differences between socioeconomic classes in this sample. It also means that education might not have as much of an influence as first assumed.

However, education showed significant differences between categories when examining uncertainty. One reason why higher education without a degree was more confident than participants with bachelor's, master's, or Ph.Ds.' could be that they have completed fewer years of education. Hence, participants who have completed a degree might be more critical when evaluating content, and when crucial information like sender is missing, this could contribute to greater uncertainty.

Admittedly, one weakness of the sample concerning education is that the survey did not acquire a representative sample for each category on the education variable. In the sample, only 8 participants reported only to have completed primary and secondary school. Moreover, there were only 37 participants who reported to have completed a Ph.D. and to have higher education with no degree. 238 respondents reported to have completed a bachelor's degree, and 194 participants had a master's degree. Hence, the other categories are underrepresented, and this might have influenced the results. Perhaps, if the sample had been more representative concerning the educational background, the variable would have shown significant influence.

6.2.3 Digital literacy

H3 stated that "*Respondents who report high digital literacy will perform better when classifying deepfakes than not digitally literate respondents*." This is because becoming digitally literate might have taught about what cues to look for when evaluating the authenticity of media content. Moreover, being digitally literate increases the chances of previously being exposed to digital content of different types. However, results from the analysis show no correlation between digital literacy and the ability to detect deepfakes. Moreover, it also shows no correlation between digital literacy and the number of uncertain responses, meaning that, for this sample, being digitally literate did not affect their ability to detect deepfakes.

However, because digital literacy did not significantly influence the improved model, H3 can be discarded.

Digital literacy is a variable that might contain measuring errors since it asks about subjective perceptions. Although respondents were given a definition of the term and what it includes, their perceptions might lead to differences in their responses. This might be one reason why the variable showed no significant effect. Further, another affecting factor might be the choice of method. Considering that the survey was made and distributed online, it automatically excludes people who have low digital literacy since they most likely would not find or complete the survey. This could have affected the results since some groups of the target audience were left out.

Another point that could influence the results is that in the term digital literacy, being able to evaluate information and sources are considered highly important. However, in the survey, respondent was not given a sender as they would have in real life, and the content of the video was kept at a minimum with only 10 seconds per video. Hence, respondents were not able to critically evaluate where the information came from. Perhaps if this kind of information were provided, digital literacy would significantly affect the number of correct answers.

Another interesting point is that digital literacy is significant in the simple models of both the number of correct answers and the number of uncertain answers. The variable, however, becomes insignificant when control variables were added in the improved models. This is also the case when examining uncertainty, where there was a negative significant correlation in the simple model, but this became insignificant in the adjusted model. This might mean that there are some influences, but that the influence is not significant enough.

Consequently, because the variable became insignificant in the improved models, it did not influence participants' performance or confidence. This is contrary to previous research showing that previous exposure to digital content positively affected participants' abilities to detect fake digital content (Schetinger et al., 2017; Fan et al., 2012). This is a sign that

classifying deepfakes is not considered when evaluating whether someone is digitally literate. Perhaps, more structurally, the term should cover more aspects according to the needs and threats of tomorrow. Lastly, there was no interaction between digital literacy and higher education, which is an interesting finding because of how the ability to evaluate sources and information critically is considered an essential knowledge exchange from higher education, as well as a crucial quality to being digitally literate (NOU 2019: 2, 2019; Sande, 1989).

6.2.4 Trust in news

H4 stated that "*There is a negative relationship between having increased trust in news outlets and the ability to identify deepfakes correctly*." Trust in news is a complex variable to measure correctly, as it asks for subjective perceptions. This might lead to different people perceiving the question differently. Although the variable did not show a significant influence, the sign of the coefficient when looking at uncertainty is negative. This tendency coincides with previous research (Vaccari and Chadwick, 2020). However, because the variable was not significant in any of the models, H4 can be discarded.

Admittedly, I did not manage to reach out to a representative amount of people who consider themselves to have very low trust or low trust. According to Newman et al. (2018), Norwegians, in general, have low trust in news, which means that there is a significant group of the Norwegian population that is not represented in the sample. Consequently, this might have influenced the results and might have caused the variable to be insignificant in the regression models. However, even though the variable was not statistically significant, it has a positive sign concerning the number of correct responses, which is the opposite of what H4 stated.

This low trust in news outlets might affect their response to what is published in the real world, but in the context of this research, trust in news outlets showed no significant effect on participants' ability to detect deepfakes. Further, it also showed no effect on participants' confidence by looking at the number of uncertain answers. This stands in contrast to previous research conducted on the topic of classifying deepfakes, where participants who gave uncertain answers showed significantly lower levels of trust in news, particularly on social media (Vaccari and Chadwick, 2020). Such low levels of trust in news outlets might increase knowledge gaps, because people tend to trust the opinions of family and friends more than commercial actors (Aalen, 2016). Hence, if trust in news outlets is low, but trust in friends and family is high, people might not be able to critically evaluate the information they are fed, and they might miss out on essential information which could increase knowledge gaps.

6.2.5 Previous knowledge of the subject

Having previous knowledge of the subject was researched in the logistic model in the analysis. H5 stated a negative correlation between having previous knowledge of the subject in the video and giving an uncertain answer. This seems to be confirmed in 6 of the videos, portraying River Phoenix, Zendaya, Steve Carell, Margot Robbie, Edward Snowden, Melania Trump, where previous knowledge of the subject has a significant negative influence on giving an uncertain answer. Giving uncertain answers on the videos of Kim Jung-Un, Nicolas Cage, and Mette Frederiksen is also correlated with previous knowledge of the subject but was less significant than the other videos. This is in line with previous research (Khodabakhsh et al., 2019)

When researching whether previous knowledge influenced the number of correct responses when classifying, the analysis showed that it did affect some videos, but not all. It did influence the classification of the videos portraying Zendaya, Steve Carell, Dua Lipa, Margot Robbie, Kim Jung-Un, Melania Trump, and Nicolas Cage, showing a significant positive correlation. Ergo, for 7 of the 16 videos, having previous knowledge of the subject positively affected participants' ability to identify them correctly. However, for the remaining nine videos, having previous knowledge did not affect their ability to identify the videos correctly. Because of this, I will argue that previous knowledge of the subject, in general, does not correlate with the ability to identify deepfakes correctly.

However, because of the significant influence in giving uncertain answers in six of the videos and tendencies for influence in another three, I will argue that H5 can be retained, and that previous knowledge of the subject decreases the odds of giving an uncertain answer.

These results could be linked to confirmation bias proclaims that people tend to accept the message if the content seems familiar (Brinsky, 2015; Newman et al., 2015). Hence, familiarity with the subject might influence whether participants accept the message and might lead to less uncertainty when classifying videos. Moreover, people tend to remember better the content they were exposed to through audiovisual media than text (Vaccari and Chadwick, 2020). Hence, if participants have previously been exposed to an audiovisual representation of a subject portrayed in the video, they might better remember their traits, leading to less uncertainty when classifying.

6.2.6 Other factors

Additionally, control variables were added to ensure a fuller understanding of what might affect Norwegian's abilities to detect deepfakes. These control variables were gender, expected performance, political interest, internet use per day, and previous knowledge of the video in question.

Gender showed no significant effect on either dependent variable, meaning that men and women detect deepfakes with an equal success rate in this sample. This is, to some extent, contrary to the findings of Nightingale et al. in 2017. However, gender does not seem to influence the ability to detect fake content or is not an included variable in other previous research projects (Khodabakhsh et al., 2019; Fan et al., 2012; Zheng et al., 2019; Vaccari and Chadwick, 2020). Hence, it seems like the gender perspective is not as important when looking at knowledge gaps generally, especially deepfake detection.

Expected performance positively influenced the ability to correctly classify videos, which means that respondents who reported increased confidence in performance also performed better. This is also in line with what Khodabakhsh et al. (2019) found in their research. Expected performance also affected the number of uncertain answers, however, negatively. As such, participants who expected to perform well were more confident when distinguishing, hence having fewer uncertain answers. Some participants might expect to perform better than others because of their perception of their own digital literacy being higher than others or increased previous knowledge of deepfakes.

Moreover, since expected performance in many ways measures confidence, we could also have expected to see an interaction between expected performance and gender (Cho, 2017). However, no interaction was found.

These are interesting findings considering expected performance, put in other words, asks about participants' confidence when classifying videos. It is also interesting to see the positive correlation, considering previous research has shown that people tend to overestimate their abilities to distinguish real content from fake (Thaler, 2019). Further, according to Newman et al. (2018), Norwegians also reported very low overall confidence concerning the detection of fake media compared to other countries. This could indicate that confidence in performance does improve performance, but that performance is still worse than what participants expected. One reason why some participants report higher levels of expected performance than others might be that they have more experience and knowledge of the term than those who are learning about deepfakes for the first time. Therefore, one reason why expected performance correlated with performance could be that the respondents who expected to perform good had more experience and knowledge about deepfakes in general, and perhaps knew what to look for.

However, expected performance is another subjective variable that may contain measuring errors. Since respondents were asked how well they expected to perform without any indication of what a "good" performance would be, respondents might have perceived the question differently. Moreover, the survey only had 6 respondents who reported their expected performance as "very bad." Therefore, the survey seems to have failed to reach the participants who would have reported low confidence in performance and failed to create a more nuanced reflection of the low confidence, as seen in Newman et al. (2018). This and the potential measuring error of it being a subjective question might have influenced the results.

One reason why there were such few respondents with low expectations of their performance might be that they left the survey before completing it or never entered the survey in the first place. In a comment on the Facebook post, this was pointed out that having the word "deepfake" in the title might have scared people away from entering. Perhaps people who might have expected to perform poorly are the same people who refrained from entering the survey because they did not know what a deepfake was. This is in line with what Sande (1989) argued concerning knowledge gaps. People who already know tend to attain knowledge with more ease than those who do not have the required knowledge, who might refrain from participating in unfamiliar exercises.

This became quite apparent when Facebook users were commenting on the Facebook post that they chose not to complete the survey because they perceived it as "silly" or too hard to evaluate without more context. These comments show that the results might be affected by lacking participants who might have performed poorly, since they did not complete the survey. Hence, the results showing that people, in general, classify with a 57.3% success rate might be overestimated. Simultaneously, these comments on the Facebook post might also have caused other potential respondents to avoid completing the survey because of the negative response by others. I did not delete the comments or close the comment section, hoping that potential participants would also read my responses and get a better understanding of the survey.

Political interest showed no significant influence on neither ability to detect deepfakes nor confidence when doing so. However, the sign of the coefficient is negative in the detection of deepfakes and positive in uncertainty when classifying. Hence, when detecting deepfakes, having less political interest would have been advantageous if the variable would have been significant. Having increased interest in politics would also have increased

uncertain answers. This is interesting because it is contrary to earlier beliefs of correlation. However, the correlations are not significant. Hence, the assumed correlation between political interest and knowledge of politicians might not be as strong and influencing when it comes to detecting deepfakes. One reason for this could be that even though respondents could have prior knowledge of several politicians, they are most often portrayed in newspapers using images, instead of clips on TV. Hence, respondents might have extensive knowledge about the politician, but not as much of their facial muscle contractions or head movements. Therefore, they might still struggle with evaluating a video portraying them.

Internet use showed a curvilinear correlation with correctly classifying deepfakes with an inverted U-shape. Consequently, the ability to detect deepfakes increases with the number of hours spent on the internet and then plateaus and might decrease. Looking at attachment 3, we see those participants in category 4, spending 9-11 hours on the internet per day, surpassed the other groups when detecting deepfakes. This is contrary to former beliefs that internet use would have a positive linear correlation with the ability to detect deepfakes. This was based on previous research saying that prior exposure to digital content increases the ability to detect fake digital content (Schetinger et al., 2017).

Internet use also significantly influences the number of uncertain answers, also here being curvilinear; however, the correlation results in a U-shape. Hence, increased internet use decreases the number of uncertain answers until category 4, representing 9-11 hours, where it slightly increases the number of uncertain answers and then plateaus. This correlation can be seen in attachment 4.

One possible reason for these curvilinear correlations could be that participants who spend more than 9-11 hours on the internet per day, are doing something more time-consuming than browsing the internet. These time-consuming activities could for instance be to play videogames or watch tv shows for extended periods of time. Therefore, participants could be spending more than 12 hours on the internet per se but taking part in activities that does not expose them to fake news or improve their digital literacy. This could be a reason why those spending 9-11 hours performed better and were less uncertain than those spending more time on the internet.

However, hours spent on the internet solemnly relies on self-reporting. Previous research of self-reporting of internet use has shown that the accuracy of self-reporting compared to the actual time spent tends to be quite low (Scharkow, 2016). Respondents tend to overreport, which could have influenced the results of this study. Hence, participants might be spending less time on the internet than what they report, and their overall exposure to digital content might not be as high as they think.

One point of discussion that should be considered is that due to the COVID-19 pandemic, most parts of people's lives have become digital, forcing people to spend more time on the internet. Hence, many hours spent on the internet per day do not necessarily mean that respondents are surfing the internet, exposing themselves to different media content. They are most likely also spending time on the internet working from home, watching digital lectures, or doing homework. Hence, this variable and its influence might not accurately represent how internet use affects the ability to detect deepfakes. On a positive note, however, this increased in time spent on the internet might contribute to evening out potential gaps in digital literacy as it has forced people to learn new tools.

Further, previous knowledge of the video was added as a control variable in the logistic regression model. This variable showed a significant correlation with correctly classifying 7 videos, portraying Stefan Löfven, Steve Carrell, Margot Robbie, Edward Snowden, Kim Jong-Un, Nicolas Cage, and Mette Frederiksen. This is contrary to previous beliefs that having previous knowledge of the video or other versions would increase the odds of correctly classifying it since this was only the case for 7 out of the 16 videos. All the correlations were positive, except for Mette Frederiksen, where having previous knowledge of the video decreased the odds of correctly classifying it.

However, previous knowledge of the video in question showed significant correlations with giving an uncertain answer. In 12 of the 16 videos, having previous knowledge of the video led to a decrease in the odds of giving an uncertain answer, where 5 of the videos showed perfect collinearity. Hence, no one who had previously seen the video gave an uncertain answer. This is in line with previous beliefs that previous knowledge of the video would increase participants' confidence when answering. This, again, might be caused by the "truthiness effect," which says that people tend to accept a message if the subject seems familiar (Brinsky, 2015; Newman et al., 2015).

One peculiar finding was that there seemed to be a negative correlation between having previously seen the video of Mette Frederiksen and correctly classifying it as a deepfake. One reason for this could be the "illusory truth effect." The deepfake portraying the Danish prime minister was featured in the news, and it might also have appeared on Norwegian's radar (Kott, 2021). Because of this recognition, participants might have seen it and remembered it as true, despite the indication that it was fake. (Franks and Waldman, 2019; Aumyo and Barber, 2021). Instances like these might contribute to growing knowledge gaps, as it influences people's ability to remember what media is fake and real.

Moreover, several videos showed perfect collinearity when examining the influence of previous knowledge of the video. One highly influencing aspect of this perfect collinearity could be that a very small number of participants reported to have seen the videos in these cases. With numbers ranging from 3 to 26, chances are high that all those respondents were able to correctly classify the videos. This is not to say that previous knowledge of the video does not significantly influence the odds of giving an uncertain answer. The perfect correlation is correct but might give the wrong impression of the effect of having previously seen the video. However, since the variable did show significant influence in 12 videos, having previous knowledge of the video can, nonetheless, be considered an influencing factor when examining uncertainty in respondents.

Lastly, respondents were asked about which cues were most important when classifying the videos. This question asks about properties of the video instead of the respondents' properties, as most other questions did. The analysis showed that "unsure" was a quite common response for both real videos and deepfakes. Further, there were substantial differences between which cues were chosen for real videos and deepfakes. Head/face, movements, sound/voice, and synchronicity in sound/voice were most common for deepfakes. In contrast, shadows, synchronicity in movements, background, and the subject's body were most used in real videos. These results match, to some extent, previous research conducted by Khodabakhsh et al. (2019) and Fan et al. (2012) who also found that the most used cues were in the head and face area and lighting.

Additionally, from the written responses, I saw that respondents sometimes used incorrect or misleading cues to classify the videos. One example of this is the video of Kim Jong-Un mentioned above, where respondents based their decision on their belief that he could not speak English when he went to school in Europe and was taught English from early on. Similarly, this was also a cue used in the real video of Angela Merkel, where respondents were arguing that she never uses English when speaking. There are also similar examples for every video, where respondents have chosen an unreasonable and incorrect cue when deciding. This is also what Nightingale et al. (2017) found in their research.

There was also quite a big difference in which cues were used to classify real videos, and which were used on deepfakes. One reason might be that the cues might be more apparent when it comes to deepfakes. Although the quality of today's deepfakes is quite high, perfect synchronicity is still hard to achieve. However, in the real videos, other cues might be used because there is perfect synchronicity in the voice and movements. Moreover, because the lighting is natural, the shadows are also correct in relation to their face and movements.

These cues are interesting because synchronicity in sound/voice, synchronicity in movements, and blinking are difficult to perfect when creating deepfakes and are, therefore, cues that should be used to recognize a deepfake (Tolosana et al., 2020; Mirsky and Lee, 2021). Shadows are also an important cue to look for when distinguishing between real and fake videos since correct lighting and shadowing is hard to achieve when creating a deepfake (Westerlund, 2019; Nightingale et al., 2017). Hence, I would argue that Norwegians, in general, use some of the most important cues when it comes to correctly classifying videos, but sometimes also depend on incorrect and misleading cues that might deceive them in real life. This, in turn, might lead to increased knowledge gaps in society.

6.3 Summarizing discussion and reflections

In this sub-chapter, I will summarize the discussion and reflect around the relevance of these findings in a broader context.

To summarize, the sample showed an overall success rate of 57.3% when classifying videos. Age influenced both performance and uncertainty, whereas education only influenced uncertainty. Previous knowledge of the subject and video both influenced uncertainty, equivalent to expected performance. Lastly, internet use had a curvilinear influence on both performance and uncertainty, but with opposite directions.

When studying the research question, I have considered the properties of the participants and the properties of the videos. I argue that the properties of the video also highly influence whether people can recognize them as fake. However, the results should be seen in the context in which the data was gathered. Hence, there is reason to believe that Norwegians would surpass their performance in the survey in real life when the content most likely will be more dramatic, and they can evaluate the sender's authenticity.

Further, the feedback questions showed that the survey contributed to increase the awareness of what deepfakes are. Another mentioned that they became increasingly aware of which cues might be used to detect deepfakes as they completed the survey. Based on the categories of the question asking which cues were most important when deciding, the participant took notice and used them when classifying the upcoming videos. Moreover,

most feedback revolved around wrote that the survey was an eye-opener, and that detecting deepfakes was considerably harder than they expected it to be.

Because of this, I believe that the result from this study is an important contribution to the research field for several reasons. Firstly, a Norwegian sample has only been the target of research in smaller projects. Because the Norwegian population tends to stand out compared to other European countries in other contexts, the data collected from this survey is a valuable resource to understand further differences in populations, particularly in the context of deepfakes. Moreover, as mentioned in the feedback, the survey was an eye-opener for quite a few respondents, and contributed to increased awareness of the phenomenon. Lastly, the research has contributed to another comprehensive study examining the detection of the phenomenon of deepfakes, with a detailed theoretical framework from a variety of research traditions.

6.4 Limitations

One recurring limitation of the study is that some groups of the target audience are underrepresented. Although age groups and gender were somewhat correctly represented compared to their share of the population, the sample lacked enough representatives from all educational backgrounds, trust in news outlets, digital literacy, and expected performance.

Another weakness was the inclusion of the "other" option when mentioning important cues used in the classification of the videos. As it turned out, most participants who ticked the "other" box and wrote their own responses could, in most cases, have chosen one of the other categories already mentioned. A written response like "choice of words" could simply have gone under the category "Words/speech" that was already mentioned as an alternative. Making an "other"-box lowered the threshold for respondents to use this instead of trying to fit their answers into one or more of the already given categories. This led to approximately 40-50 of such written responses per question. Simultaneously, as these written responses became string variables, it was impossible to include the written responses in the regression analysis because of the lack of resources to manually transform them to numerical values. Hence, these responses were not included in the analysis, and a considerable number of clues were therefore excluded. This, in turn, could have influenced the results. Instead of adding an "other"-box with the possibility of a written response, I should have communicated more clearly that respondents should try to the best of their abilities to fit their cues into the already given categories.

However, despite the weaknesses, I will argue that this research is still highly important in the field of deepfake detection, mis- and disinformation. With a high number of observations on a previously limited researched population, the research contributes to an increased understanding of how Norwegians consider, treat, and evaluate fake content. I argue that the result from this research is not necessarily only confined to this sample. The new dataset of deepfake detection can also lay the foundation for further data analysis and inspire similar, but more intricate research projects.

6.5 Outlook: Suggested responses to deepfakes

As this research has shown, the survey respondents are not particularly trustworthy when classifying deepfakes, with a success rate of 57.3%. Hence, I would like to use the last part of this discussion to shortly introduce other ways of detecting deepfakes. Several

research fields are working to counteract the use of deepfakes to spread disinformation. Evidently, it seems like digital literacy might not be enough to avoid being misled by the deepfakes of tomorrow. Hence, other measures must be taken to avoid the use of deepfakes in the spread of misinformation.

Firstly, numerous software has been made in the past few years to recognize deepfakes based on asynchrony in facial movements, blinking, lips, and other features (Kaur et al., 2020; Kim et al., 2019; Đorđević et al., 2019; Kumar et al., 2020). This kind of technology not only attempts to detect deepfakes but can contribute to authenticating and also preventing that content be used to make deepfakes (Westerlund, 2019). However, although software for deepfake-detection is improving daily, the problem with such AI-based technology is that malicious actors will often use them to improve their software for making deepfakes. This makes them harder to detect and makes the deepfake-detection software increasingly ineffective (Vaccari and Chadwick, 2020; Chesney and Citron, 2019). These kinds of deepfake-detection software are also dependent on large and diverse data sets as well as constant updates to be able to detect deepfakes, and even with this, the software might not be able to detect deepfakes with 100% accuracy (Hussain et al., 2021; Chesney and Citron, 2019).

Blockchain is another technology that might be an alternative. The thought behind using Blockchain is to ensure Proof of authenticity, which is considered to be a critical point in the battle towards manipulated content (Hasan and Salah, 2019; Donovan and Paris, 2019). By using so-called smart contracts, blockchain will allow internet users to trace content back to its original state and hence be able to see whether the content has been tampered with. The code needed to make smart contracts is also publicly available on Github (Hasan and Salah, 2019).

Others consider legal solutions to solve the problem by introducing criminal laws and administrative action to tackle revenge porn and political smear campaigns (Meskys et al., 2020). However, these intentions will most likely not stop malicious producers of deepfakes from making and spreading them (Coldewey, 2019). Donovan (2020) called for social media companies to take responsibility and become more transparent in their detection and flagging of fake news. One response to the growing challenge with the spread of deepfakes was Facebook, in collaboration with Amazon Web Services, Microsoft, and others, launched in 2019 a "Deep Fake Detection Challenge." People who develop open-source code or produce research to identify deepfakes could there be granted up to 10 million dollars and other awards (Jenkins, 2020; Knight, 2020; Skibba, 2020). Facebook also banned deepfakes on its platform as an attempt to stop the spread of misinformation on its page. The challenge attracted more than 2000 participants, and more than 35000 detection models emerged from it (Jenkins, 2020).

In conclusion, the possibilities are many, and the methods are getting more sophisticated, proportionally with the sophistication of deepfake-technology. A 100% accurate detection software seems almost impossible to develop, and imposing smart contracts seems to be too invasive to be realistic. Hence, it is still crucial to understand how people are affected by such fake content and teach them what to look for when evaluating the authenticity digital content. Overall, an increasing awareness of deepfakes existence might be necessary, so that people are aware and attentive when exposed to them. This is something I hope my thesis have contributed to.

7 Conclusion and future research

In this research project, I have investigated the research question, "To what extent are Norwegians able to recognize deepfakes and which factors affect this"? The research was conducted using a quantitative online survey asking participants to classify 16 videos, where 8 were fake, and 8 were genuine. The survey has 682 responses and therefore presents the most extensive study on deepfake detection in the Norwegian context to date.

The main findings of the research is that participants had an overall success rate of 57.3% when classifying videos. Separating real and fake videos showed that the success rate was 66.04% for real videos and 48.59% for deepfakes. This is a quite low success rate, slightly higher than the success rate of randomly guessing.

Further, older participants performed worse than younger participants, and gave more uncertain answers. This could be caused by a digital divide between age groups. Having previous knowledge of the subject of the video decreased the odds of giving an uncertain answer, which might be caused by confirmation bias and the truthiness effect. Expecting to perform well had a positive influence on overall performance and confidence when responding. This might mean that respondents who expected to perform well had more extensive knowledge and experience with deepfakes prior to completing the survey.

The number of hours spent on the internet influenced performance when detecting in a curvilinear way, where respondents spending 9-11 hours on the internet per day surpassed other groups. Internet use also had a curvilinear correlation with uncertain answers, where respondents spending 9-11 hours on the internet gave the lowest number of uncertain answers. These correlations might be caused by the fact that participants spending more than 9-11 hours on the internet are doing activities that does not expose them to fake news or increase their digital literacy to a great extent, like playing video games or watching tv-shows.

Overall, the ability to classify videos in the population is relatively low. I want to highlight here that only two respondents were able to classify all videos correctly. These results coincide with Norwegian's previous reporting of low confidence levels regarding their ability to detect fake media in real life. This, in turn, might be caused by the relatively low exposure of such fake content and there being few examples of deepfakes used in real-life contexts as disinformation. Most instances where deepfakes have been used for spreading disinformation have come from countries other than Norway (Elliott, 2021; Harwell and Okazaki, 2021; Donovan and Paris, 2019; Van Boom, 2019). Simultaneously, Norway has tried to combat the spread of mis- and disinformation by using fact-checking services such as faktisk.no, but because of the illusory truth effect, this might contribute to spreading even more misinformation. However, Norwegian's decreased ability to detect deepfakes might cause problems if they infiltrate our lives to a greater extent in the future. Hopefully, by then, there are powerful and highly successful software programs identifying the fake content for us since there might be a lot at stake, and we cannot rely on people to evaluate on their own.

For 2 out of 5 hypotheses, concerning age and previous knowledge of the subject, there is conformity with previous research on human detection of fake news. However, research on human detection of deepfakes, is scarce. In spite of the lack of representation from

certain groups in the target population, I argue that the result from this research contributes to a better understanding of what actually influences peoples' abilities to detect deepfakes. Moreover, because of the large number of observations with a diverse and representative age distribution, I argue that the result of this research is not necessarily only applicable to this sample.

Suggestions for future research include further examination of the factors influencing the ability to detect deepfakes, using more complex analyses. This dataset allows for numerous analyses that go beyond the time and resources available for this project and the scope of this thesis does not allow me to examine the data with the desired depth. However, making the data available online for further research might facilitate more and increasingly detailed research.

Additionally, further suggestions for research projects include a higher number of deepfake videos of different qualities. Moreover, future research examining whether participants are better at detecting deepfakes portraying a subject of their own ethnicity would be valuable. Experimental research in controlled environments could also be valuable contributions to the field, as well as an increased focus on the development of deepfake-detection software. With the previous research on human detection in mind, it seems that detection software might be the most effective way to prevent the spread of mis- and disinformation through deepfakes.

Referanser

AGARWAL, S., FARID, H., GU, Y., HE, M., NAGANO, K. & LI, H. 2019. Protecting World Leaders Against Deep Fakes. *Conference on Computer Vision and Pattern Recognition.* Long Beach, California.

ALDWAIRI, M. & ALWAHEDI, A. 2018. Detecting Fake News in Social Media Networks *Procedia Computer Science,* 141**,** 215-222.

ALEXANDROU, A. & MARAS, M.-H. 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof,* 23**,** 255-262.

ALNES, E. 2021. *Cheerleader-mor skulda for å ha starta svertekampanje på nett* [Online]. NRK. Available: https://www.nrk.no/urix/mor-skulda-for-a-lage-falske-videoar-for-a-sverte-dotteras-cheerleader-rivalar-1.15418815 [Accessed 13.04 2021].

ANTINORI, A. 2019. Terrorism and DeepFake: from Hybrid Warfare to Post-Truth Warfare in a Hybrid World. *ECIAIR 2019 European Conference on the Impact of Artificial Intellegence and Robotics.*: Academic Conferences and Publishing Limited.

AUMYO, H. & BARBER, S. J. 2021. The effects of repetition frequency on the illusory truth effect: Principles and Implications. *Cognitive Research,* 6.

AYYUB, R. 2018. *I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me* [Online]. The Huffington Post. Available: https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316 [Accessed 15.04 2021].

BATES, M. E. 2018. Say What? "Deepfakes" Are Deeply Concerning. *Online Searcher,* 42**,** 64.

BBC. 2019. *The Queen's Christmas Broadcast 2019* [Online]. YouTube. Available: https://www.youtube.com/watch?v=KgvZnxNAThM [Accessed 28.05 2021].

BETHLEHEM, J. 2010. Selection Bias in Web Surveys. *International Statistical Review,* 78**,** 161-188.

BHUTTA, C. B. 2012. Not by the Book: Facebook as a Sampling Frame. *Sociological Methods & Research,* 41**,** 57-88.

BOCZKOWSKI, P. & LIEVROUW, L. A. 2008. Bridging STS and communication studies: Scholarship on media and information technologies. . *The Handbook of Science and Technology Studies* 949-977.

BRANDTZÆG, P. B., HEIM, J. & KARAHASANOVIĆ, A. 2011. Understanding the new digital divide—A typology of Internet users in Europe. *International Journal of Human-Computer Studies,* 69**,** 123-138.

BRINSKY, A. J. 2015. Rumors and Health Care Reform: Experiments in Political Misinformation. *British Journal of Political Science,* 47.

BUZZFEEDVIDEO. 2018. *You Won't Believe What Obama Says In This Video!* [Online]. Youtube.com. Available: https://www.youtube.com/watch?v=cQ54GDm1eL0 [Accessed 13.04 2021].

CHANNEL 4. 2020. *Deepfake Queen: 2020 Alternative Christmas Message* [Online]. YouTube. Available: https://www.youtube.com/watch?v=IvY-Abd2FfM [Accessed 28.05 2021].

CHESNEY, R. & CITRON, D. K. 2019. 21st Century-Style Truth Decay: Deep Fakes and the Challenges for Pivacy, Free Expression, and National Security. *Maryland Law Review* 78.

CHETTY, K., QIGUI, L., GCORA, N., JOSIE, J., WENWEI, L. & FANG, C. 2018. Bridging the digital divide: measuring digital literacy. *Economics: The Open-Access, Open-Assessment E-Journal,* 12**,** 1-20.

CHO, S.-Y. 2017. Explaining Gender Differences in Confidence and Overconfidence in Math *SSRN Electronic Journal*.

CITRON, D. & CHESNEY, B. 2019. Deep Fakes: A looming Challenge for Privacy, Democracy, and National Security. *California Law Review***,** 1753-1820.

COLDEWEY, D. 2019. DEEPFAKES Accountability Act would impose unenforceable rules — but it's a start. Available: https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAANRfDI5VxarhFGUxbMUqIHu2EQzTOuJTpa5yOecI5tBXWT82Y5dFoZuWnt1bIu3SPV3V6oy5ACJU_gv5AA2tzP7Z6VNh1vc6oFESos1syLwbdLAafIlL6c113mYeL8eMawUL2SNL5aKPMYKlISfP5BKbmq0d8Q7Ung33F9-s2mIs.

COUPER, M. P., KAPTEYN, A., SCHONLAU, M. & WINTER, J. 2007. Nonconverage and nonresponse in an Internet survey. *Social Science Research,* 36**,** 131-148.

DAY, C. 2019. The Future of Misinformation. *Computing in Science & Engineering***,** 108.

DIAKOPOULOS, N. & JOHNSON, D. 2020. Anticipating and addressing the ethical implications of deepfakes in the context of election. *New media & society***,** 1-27.

DIGCOMP 2.0. 2019. *The Digital Competence Framework 2.0* [Online]. EU Science Hub. Available: https://ec.europa.eu/jrc/en/digcomp/digital-competence-framework [Accessed 19.04 2021].

DONOHUE, G. A., TICHENOR, P. J. & OLIEN, C. N. 1975. Mass Media and the Knowledge Gap: A Hypothesis Reconsidered. *Communication Research,* 2**,** 3-23.

DONOVAN, J. & PARIS, B. 2019. Deepfakes and Cheap Fakes. The manipulation of audio and visual evidence. *Data & Society***,** 0-47.

ĐORĐEVIĆ, M., MILIVOJEVIĆ, M. & GAVROVSKA, A. 2019. DeepFake Video Analysis using SIFT Features. *27th Telecommunications forum TELFOR 2019.* Belgrade, Serbia.

ELLIOTT, J. K. 2021. *Cheerleader's mom charged in deepfake plot against rival teens* [Online]. Global News. Available: https://globalnews.ca/news/7697534/mom-cheerleader-deepfake-nudes-daughter/ [Accessed 13.04 2021].

ERSTAD, O. 2006. A new direction? Digital literacy, student participation and curriculum reform in Norway. *Education and Information Technologies,* 11**,** 516-429.

EVNAS, J. R. & MATHUS, A. 2005. The Value of Online Surveys. *Internet Research* 15**,** 195-219.

FAN, S., NG, T.-T., HERBERG, J. S., KOENIG, B. L. & XIN, S. 2012. Real or Fake? Human Judgements about Photographs and Computer-generated Images and Faces *SIGGRAPH Asia 2012 Technical Briefs.* Singapore.

FARID, H. & BRAVO, M. J. 2012. Perceptual discrimination of computer generated and photographic faces *Digital Investigation,* 8**,** 226-235.

FERNANDO, T., FOOKES, C., DENMAN, S. & SRIDHARAN, S. 2019. Exploiting Human Social Cognition for the Detection of Fake and Fraudulent Faces via Memory Networks.

FJØRTOFT, T. O. 2017. *Norge i Europatoppen på digitale ferdigheter* [Online]. Statistisk sentralbyrå. Available: https://www.ssb.no/teknologi-og-innovasjon/artikler-og-publikasjoner/norge-i-europatoppen-pa-digitale-ferdigheter [Accessed 10.05 2021].

FLATT, C. & JACOBS, R. 2019. Principle Assumptions of Regression Analysis: Testing, Techniques, and Statistical Reporting of Imperfect Data Sets. *Advances in Developing Human Resources* 21**,** 484-502.

FRANKS, M. A. & WALDMAN, A. E. 2019. Sex, lies and videotape: Deep fakes and free speech delusions. *Maryland Law Review* 78**,** 892-898.

FREILING, I., KRAUSE, N. M. & SCHEUFELE, D. A. 2021. Believing and sharing misinformation, fact-checks, and accurate information on social media: The role of anxiety during COVID-19. *New Media & Society*.

GEROSA, T., GUI, M., HARGITTAI, E. & NGUYEN, M. H. 2021. (Mis)informed During COVID-19: How Education Level and Information Sources Contribute to Knowledge Gaps. . *International Journal of Communication,* 15**,** 2196-2217.

GROVES, R. M. 2011. Three Eras of Survey Research *Public Opinion Quarterly,* 75**,** 861-871.

HARWELL, D. & OKAZAKI, S. 2021. *A 'beautiful' female biker was actually a 50-year-old man using FaceApp. After he confessed, his followers liked him even more.* [Online]. The Washington Post. Available: https://www.washingtonpost.com/technology/2021/05/11/japan-biker-faceapp-soya-azusagakuyuki/ [Accessed 24.05 2021].

HASAN, H. R. & SALAH, K. 2019. Combating Deepfake Videos Using Blockchain and Smart Contracts *IEEE Access,* 7**,** 41596-41606.

HOLMES, O., BANKS, M. S. & FARID, H. 2016. Assessing and Improving the Identification of Computer-Generated Portraits *ACM Transactions on Applied Perception,* 13**,** 1-12.

HUSSAIN, S., NEEKHARA, P., JERE, M., KOUSHANFAR, F. & MCAULEY, J. 2021. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.*

HWANG, Y. & JEONG, S.-H. 2009. Revisiting the Knowledge Gap Hypothesis: A Meta-Analysis of Thirty-Five Years of Research. *Journalism & Mass Communication Quarterly,* 87**,** 513-532.

JENKINS, S. 2020. Facebook bans deepfakes. *Biometric Technology Today***,** 2-3.

KALSNES, B. 2019. *Falske Nyheter. Løgn, desinformasjon og propaganda i den digitale offentligheten,* Oslo, Cappelen Damm Akademisk.

KAUR, S., KUMAR, P. & KUMARAGURU, P. 2020. Deepfakes: temporal sequential analysis to detect face-swapped video clips using convolutional long short-term memory. *Journal of Electronic Imaging,* 29.

KHODABAKHSH, A., RAMACHANDRA, R. & BUSCH, C. 2019. Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content. *Eleventh International Conference on Quality of Multimedia Experience (QoMEX).* Berlin, Tyskland.

KHODABAKHSH, A., RAMACHANDRA, R., RAJA, K., WASNIK, P. & BUSH, C. 2018. Fake Face Detection Methods: Can They Be Generalized? *International Conference of the Biometrics Special Interest Group (BIOSIG) (2018).*

KIM, J., HAN, S. & WOO, S. S. 2019. Classifying Genuine Face images from Disguised Face Images. *IEEE International Conference on Big Data (Big Data).* Los Angeles, CA, USA.

KNIGHT, W. 2020. *Deepfakes Aren't Very Good. Nor Are the Tools to Detect Them* [Online]. Wired. Available: https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/ [Accessed 19.03 2021].

KORSHUNOV, P. & MARCEL, S. 2020. Deepfake detection: human vs. machines.

KOTT, S. 2021. *»Alle svin skal aflives«: Sagde Mette Frederiksen virkelig det?* [Online]. Jyllands-Posten. Available: https://jyllands-posten.dk/kultur/ECE12698849/alle-svin-skal-aflives-sagde-mette-frederiksen-virkelig-det/ [Accessed 31.05 2021].

KROGSRUD, V. K. & VELSAND, M. 2020. *Deepfakes - et undervisningsopplegg om kritisk medieforståelse* [Online]. Medietilsynet. Available: https://www.medietilsynet.no/barn-og-medier/deepfakes/ [Accessed 13.04 2021].

KUMAR, P., VATSA, M. & SINGH, R. Detecting Face2Face Facial Reenactment in Videos. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020. 2589-2597.

KUNNSKAPSDEPARTEMENTET 2017. Framtid, fornyelse og digitalisering. Digitaliseringsstrategi for grunnopplæringen 2017–2021.

KWOK, A. O. J. & KOH, S. G. M. 2020. Deepfake: a social construction of technology perspective *Current Issues in Tourism*.

LEWIS-BECK, M. S. 1993. *Regression Analysis*, SAGE Publications.

LUPAČ, P. 2018. *Beyond the digital divide: contextualizing the information society,* Bingley, England, Emerald Publishing

MARCHI, R. 2012. With Facebook, Blogs, and Fake News, Teens Reject Journalistic "Objectivity". *Journal of Communication Inquiry* 36**,** 246-262.

MCDOUGALL, J. 2019. Media Literacy versus Fake News: Critical Thinking, Resilience and Civic Engagement. *Media studies,* 10**,** 29-45.

MEHMETOGLU, M. & JAKOBSEN, T. G. 2017. *Applied statistics using STATA*, SAGE.

MESKYS, E., LIAUDANSKAS, A., KALPOKIENE, J. & JURCYS, P. 2020. Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law & Practice,* 15**,** 24-31.

MEURER, W. J. & TOLLES, J. 2017. Logistic Regression Diagnostics. Understanding How Well a Model Predicts Outcomes. *Clinical Review & Education,* 317**,** 1068-1069.

MIRSKY, Y. & LEE, W. 2021. The Creation and Detection of Deepfakes: A Survey. *Association for Computing Machinery,* 54.

MULLINIX, K. J., LEEPER, T. J., DRUCKMAN, J. N. & FREESE, J. 2015. The Generalizability of Survey Experiments. *Journal of Experimental Political Science* 2**,** 109-138.

MURRAY, L. 2020. *Kim Jong-Un. North Korean political official* [Online]. Britannica. Available: https://www.britannica.com/biography/Kim-Jong-Eun [Accessed 28.05 2021].

NEWMAN, E. J., GARRY, M., UNKELBACH, C. & BERNSTEIN, D. M. 2015. Truthiness and Falsiness of Trivia Claims Depend on Judgmental Contexts. *Journal of Experimental Psychology Learning Memory and Cognition* 41.

NEWMAN, N., FLETCHER, R., KALOGEROPOULOS, A., LEVY, D. A. L. & NIELSEN, R. K. 2018. Reuters Institute Digital News Report 2018. Reuters Institute for the Study of Journalism.

NIGHTINGALE, S. J., KIMBERLEY, W. A. & WATSON, D. G. 2017. Can people identify original and manipulated photos of real-world scenes? . *Cognitive Research: Principles and Implications,* 2.

NOU 2019: 2 2019. Fremtidige kompetansebehov II — Utfordringer for kompetansepolitikken. regjeringen.no: Kunnskapsdepartementet.

PEARSON, G. D. H. & KNOBLOCH-WESTERWICK, S. 2019. Is the Confirmation Bias Bubble Larger Online? Pre-Election COnfirmation Bias in Selective Exposure to Online versus Print Political Information. *Mass Communication and Society,* 22**,** 466-486.

PERSONS, T. M. 2020. Deepfakes. *In:* SCIENCE, T. A., AND ANALYTICS TEAM OF THE U.S. GOVERNMENT ACCOUNTABILITY OFFICE (ed.). USA.

PIETRASS, M. 2007. Digital Literacy Research from an International and Comparative Point of View. *Research in Comparative and International Education* 2.

RAHIM, Z. 2020. *'Deepfake' Queen delivers alternative Christmas speech, in warning about misinformation* [Online]. CNN. Available: https://edition.cnn.com/2020/12/25/uk/deepfake-queen-speech-christmas-intl-gbr/index.html [Accessed 28.05 2021].

RINGDAL, K. 2013. *Enhet og mangfold. Samfunnsvitenskapelig forskning og kvantitativ metode*, Fagbokforlaget Vigmostad & Bjørke AS.

RÖSSLER, A., COZZOLINO, D., VERDOLIVA, L., RIESS, C., THIES, J. & NIEßNER, M. 2018. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. Cornell University.

SANDE, Ø. 1989. *Nyheter, forståelse og kunnskapskløfter,* Volda, Møre og Romsdal distriktshøyskole.

SCHARKOW, M. 2016. The Accuracy of Self-Reported Internet Use - A Validation Study Using Client Log Data. *Communication Methods and Measures* 10**,** 13-27.

SCHETINGER, V., OLIVEIRA, M. M., DA SILVA, R. & CARVALHO, T. J. 2017. Humans are easily fooled by digital images. *Computer & Graphics,* 68**,** 142-151.

SCHEUFELE, D. A. & KRAUSE, N. M. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences Apr 2019,* 116**,** 7662–7669.

SCHIFF, K. J., SCHIFF, D. & BUENO, N. S. 2020. The Liar's Dividend: The Impact of Deepfakes and Fake News on Trust in Political Discourse.

SCHNEIDER, D. & HARKNETT, K. 2019. What's to Like? Facebook as a Tool for Survey Data Collection. *Sociological Methods & Research***,** 1-33.

SCHWEBS, T. & ØSTBYE, H. 2017. *Media i samfunnet*, Det Norske Samlaget.

SEO, H., BLOMBERG, M. & ALTSCHWAGER, D. 2020. Vulnerable populations and misinformation: A mixed-methods approach to underserved older adults' online information assessment. *New Media & Society*.

SHEN, C., KASRA, M., PAN, W., BASSETT, G. A., MALLOCK, Y. & O'BRIEN, J. F. 2019. Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New Media & Society,* 21**,** 438-463.

SIAPERA, E. 2017. *Understanding new media*, SAGE Publications Ltd.

SILBEY, J. & HARTZOG, W. 2019. The upside of deep fakes. *Maryland Law Review,* 78**,** 960-966.

SKIBBA, R. 2020. Accuracy Eludes Competitors in Facebook Deepfake Detection Challenge. *Engineering,* 6**,** 1339-1340.

SKOG, O.-J. 2017. *Å forklare sosiale fenomener. En regresjonsbasert tilnærming,* Oslo, Gyldendal Akademisk.

SLEEGERS, W. W. A., PROULX, T. & VAN BEEST, I. 2019. Confirmation bias and misconceptions: Pupillometric evidence for a confirmation bias in misconceptions feedback. *Biological Psychology,* 145**,** 76-83.

SPIVAK, R. 2019. "Deepfakes": The Newest Way to Commit One of the Oldest Crimes. *Georgetown Law Technology Review,* 3**,** 339-400.

SSB. 2021. *Befolkning* [Online]. Statistisk sentralbyrå. Available: https://www.ssb.no/en/folkemengde/ [Accessed 05.03 2021].

STOLTZFUS, J. C. 2011. Logistic Regression: A Brief Primer *Academic Emergency Medicine***,** 10099-1104.

STUPP, C. 2019. *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case* [Online]. The Wall Street Journal. Available: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [Accessed 15.04 2021].

SUNSTEIN, C. R. 2007. Republic.com 2.0. *Princeton University Press*.

TANKOVSKA, H. 2020. *Forecast of Facebook user numbers in Norway from 2015 to 2022* [Online]. Statista. Available: https://www.statista.com/statistics/568817/forecast-of-facebook-user-numbers-in-the-norway/ [Accessed 15.03 2021].

THALER, M. 2019. The "Fake News" Effect: An Experiment on Motivated Reasoning and Trust in News.

TOLOSANA, R., VERA-RODRIGUEZ, R., FIERREZ, J., MORALES, A. & ORTEGA-GARCIA, J. 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion***,** 131-148.

TSAI, H.-Y. S., SHILLAIR, R. & COTTEN, S. R. 2017. Social Support and "Playing Around": An Examination of How Older Adults Acquire Digital Literacy With Tablet Computers. *Journal of Applied Gerontology* 36**,** 29-55.

UBERTI, D. 2016. *The real history of fake news* [Online]. Columbia Journalism Review. Available: https://www.cjr.org/special_report/fake_news_history.php [Accessed 03.05 2021].

VACCARI, C. & CHADWICK, A. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*.

VAN BOOM, D. 2019. *These deepfakes of Bill Hader are absolutely terrifying* [Online]. CNET. Available: https://www.cnet.com/news/these-deepfakes-of-bill-hader-are-absolutely-terrifying/ [Accessed 24.05 2021].

VAN DER MEER, T. G. L. A. & HAMELEERS, M. 2020. Fighting biased news diets: Using news media literacy interventions to stimulate online cross-cutting media exposure patterns. *New Media & Society*

WARDLE, C. & DERAKHSHAN, H. 2017. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Brussel: Council of Europe.

WESTERLUND, M. 2019. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review* 9**,** 39-52.

ZHENG, L., ZHANG, Y. & THING, V. L. L. 2019. A survey on image tampering and its detection in real-world photos. *Journal of Visual Communication and Image Representation,* 58**,** 380-399.

ÖHMAM, C. 2020. Introducing the pervert's dilemma: a contribution to the critique of Deepfake Pornography. *Ethics and Information Technology,* 22**,** 133-140.

AALBERG, T. & ELVESTAD, E. 2012. *Mediesosiologi*, Det Norske Samlaget.

AALEN, I. 2016. *Sosiale medier*, Fagbokforlaget.

Appendix

Attachment 1: Introduction of the survey, with example of video setup

Informasjon og samtykke

Om prosjektet
Dette prosjektet omhandler deepfakes og utføres i forbindelse med en masteroppgave gjennomført av Regine Ringerud ved Institutt for sosiologi og statsvitenskap ved NTNU. Om du velger å delta i prosjektet, innebærer det at du fyller ut et spørreskjema. Dette vil ta deg omkring 20 minutter. Dine svar fra spørreskjemaet vil bli registrert og oppbevart elektronisk. Jeg vil kun bruke opplysningene om deg til forskningsformålet. Du som deltaker vil ikke kunne gjenkjennes i publikasjonen, og all data vil bli slettet etter prosjektets slutt, noe som etter planen vil være 7. Juni 2021. Om du er interessert i å vite hvilke videoer som var ekte og hvilke som var falske, vil dette publiseres på denne facebook-siden i april, når spørreundersøkelsen er avsluttet.
Hvor kan jeg finne ut mer?
Hvis du har spørsmål til studien, eller ønsker å benytte deg av dine rettigheter, ta kontakt med NTNU ved :

- Regine Ringerud (student), e-post: reginedr@stud.ntnu.no
- Lisa Reutter (veileder), e-post: lisa.m.reutter@ntnu.no
- Pieter de Wilde (veileder), e-post: pieter.dewilde@ntnu.no
- Vårt personvernombud: Thomas Helgesen, e-post: thomas.helgesen@ntnu.no

Med vennlig hilsen
Pieter de Wilde, Lisa Reutter og Regine Ringerud

Samtykke *

Jeg er over 18 år, jeg har mottatt og forstått informasjon om prosjektet og har fått anledning til å stille spørsmål. Jeg samtykker til å delta i undersøkelsen og at mine opplysninger behandles frem til prosjektet er avsluttet.

Dersom du ikke vil delta i undersøkelsen, kan du lukke fanen.

☐ Ja

Hvilket kjønn identifiserer du deg som? *

- ○ Kvinne
- ○ Mann
- ○ Annet / Ønsker ikke å oppgi

Hvor gammel er du? *

Velg …

Hva er din høyeste fullførte utdanning? *

- ○ Barne- og ungdomsskole
- ○ Videregående skole
- ○ Høyere utdanning - årsstudium
- ○ Høyere utdanning - bachelorgrad
- ○ Høyere utdanning - mastergrad
- ○ Høyere utdanning - Ph.D.
- ○ Høyere utdanning - uten fullført grad

Hvordan anser du selv din egen digitale kompetanse på generell basis? *

Digital kompetanse kan defineres som de ferdigheter, kunnskaper og holdninger som legges til grunn for å kunne bruke digitale medier.

- ○ Svært høy digital kompetanse
- ○ Ganske høy digital kompetanse
- ○ Moderat digital kompetanse
- ○ Lav digital kompetanse
- ○ Ingen digital kompetanse

Hvor mye tid bruker du på internett på en typisk dag, i timer? *

Dette inkluderer nettaviser, sosiale medier og annen surfing på internett.

- ○ 1-2 timer
- ○ 3-5 timer
- ○ 6-8 timer
- ○ 9-11 timer
- ○ 12-15 timer
- ○ Over 15 timer

Hvor høy tillit vil du si at du har til norske nyhetsmedier? *

På dette spørsmålet mener vi på generell basis, nettaviser og papiraviser som er anerkjent gjennom å ha redaktøransvar.

- ○ Svært høy tillit
- ○ Høy tillit
- ○ Moderat tillit
- ○ Lav tillit
- ○ Svært lav tillit

Hvor interessert vil du si at du er i politikk? *

- ○ Svært interessert
- ○ Ganske interessert
- ○ Litt interessert
- ○ Hverken eller
- ○ Litt uinteressert
- ○ Ganske uinteressert
- ○ Veldig uinteressert

Hva er deepfakes?

Deepfakes er lyd- eller visuelt materiale som er digitalt manipulert for å få det til å se ut som om en person sier eller gjør noe de ikke egentlig har sagt eller gjort. Det er mange ulike manipulasjons-verktøy som kan gå under deepfake-begrepet og som ligger på et bredt økonomisk spekter. I denne undersøkelsen vil jeg i all hovedsak ta for meg deepfakes som bruker kunstig intelligens for å manipulere ansikter i videoer.

Hvordan gjennomføres undersøkelsen?

I denne undersøkelsen vil du få vist til sammen 16 videoer på omkring 10 sekunder. Omtrent halvparten av videoene er ekte, og resten er deepfakes. Du kan se videoen på nytt så mange ganger du vil og se videoen i fullskjerm. Lyden på videoene kan variere i styrke, så vær obs på at den kan være høy. I noen tilfeller kan videoene også bruke noe tid på å laste inn. Jeg ber om at du ikke fullfører undersøkelsen mer enn én gang, slik at jeg kan få så nøyaktige forskningsresultater som mulig. Jeg anbefaler også å gjennomføre undersøkelsen på datamaskin eller nettbrett.

Hvor godt tror du selv du vil gjøre det når det kommer til å identifisere deepfakes? *

- ○ Svært bra
- ○ Bra
- ○ Moderat
- ○ Dårlig
- ○ Svært dårlig

Den aktuelle videoen

Trykk på knappen for å spille av. Svar deretter på spørsmålene om videoen.

Er den aktuelle videoen av Angela Merkel ekte eller falsk? *

- ○ Ekte og autentisk
- ○ Usikker / Vet ikke
- ○ Falsk

Har du sett den aktuelle videoen eller andre versjoner av den aktuelle videoen tidligere? *

- ○ Ja
- ○ Usikker
- ○ Nei

Har du kjenskap til Angela Merkel fra før? *

- ○ Ja
- ○ Usikker
- ○ Nei

Hva var den viktigste faktoren som lå til grunn for din beslutning? *

- [ ] Tidligere kunnskap
- [ ] Hode / Ansikt
- [ ] Øyne / Blunking
- [ ] Bakgrunn
- [ ] Kropp
- [ ] Skygger
- [ ] Bevegelser
- [ ] Ord / Tale
- [ ] Lyd / Stemme
- [ ] Synkronitet i bevegelser
- [ ] Synkronitet i lyd / Stemme
- [ ] Usikker
- [ ] Annet

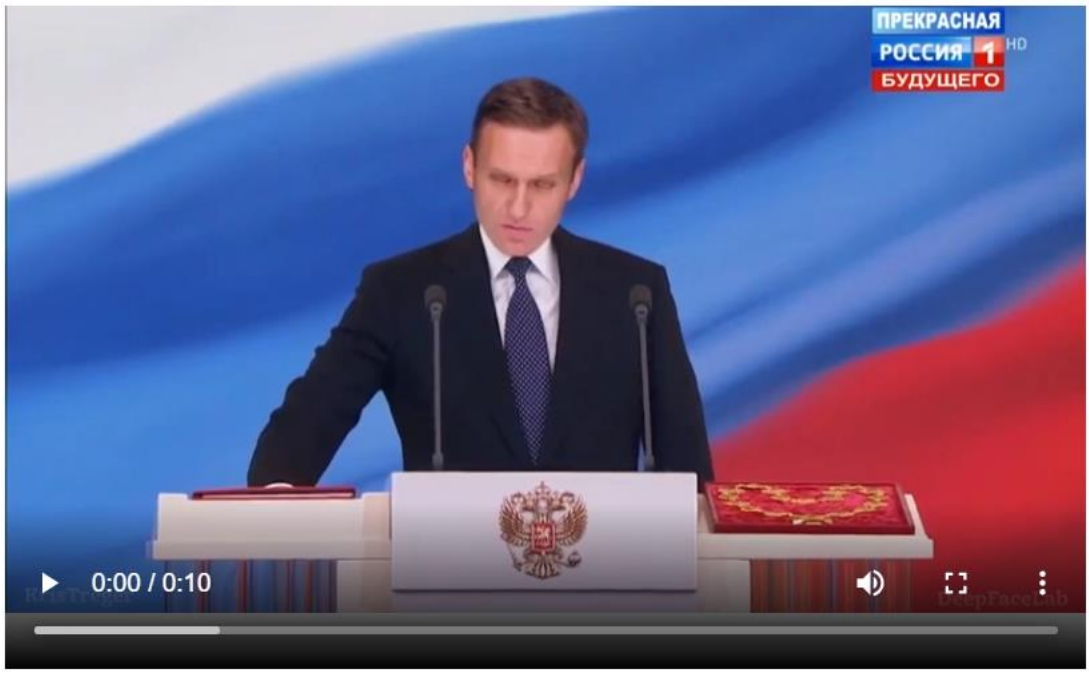
Hvilken annen faktor lå til grunn for din beslutning?

Dette elementet vises kun dersom alternativet «Annet» er valgt i spørsmålet «Hva var den viktigste faktoren som lå til grunn for din beslutning?»

Attachment 2: Screenshot, links, and explanation of the videos

Angela Merkel (genuine). Speech: "Fundamental reform of the European architecture, which will satisfy all kinds of elect or actual British wishes. I am afraid they are in for a disappointment".

Boris Johnson (genuine). Speech: "That those four cohorts, the JCVI, want to form groups of vulnerable, elderly people to get the level of immunity that they need, that's got to bed in from February".



Aleksej Navalnyj (deepfake). Speech in Russian.

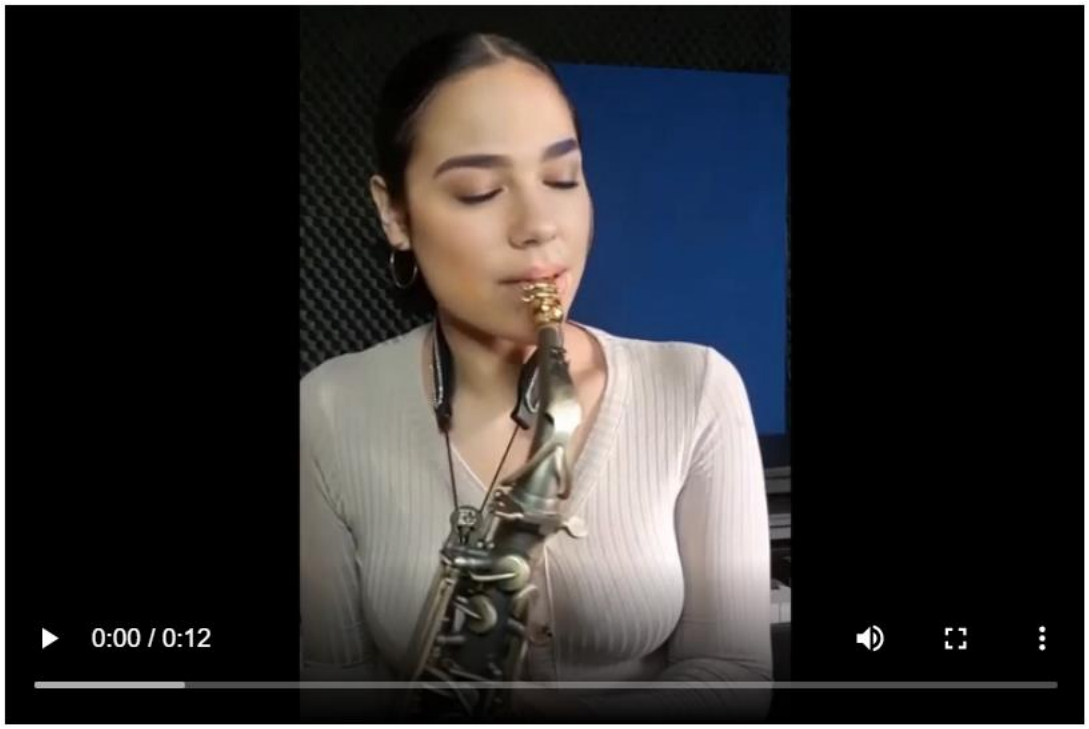
River Phoenix (deepfake). Speech: “What? Do you want to be normal; do you want to be just like everyone else? Being a freak is the best, alright, I am a freak. I have friends, Will”.

Queen Elizabeth II (deepfake). Speech: “As is so often the case, technology helped tackle the challenges we faced this year. Like many of you, when I wasn’t settling down with my husband”.

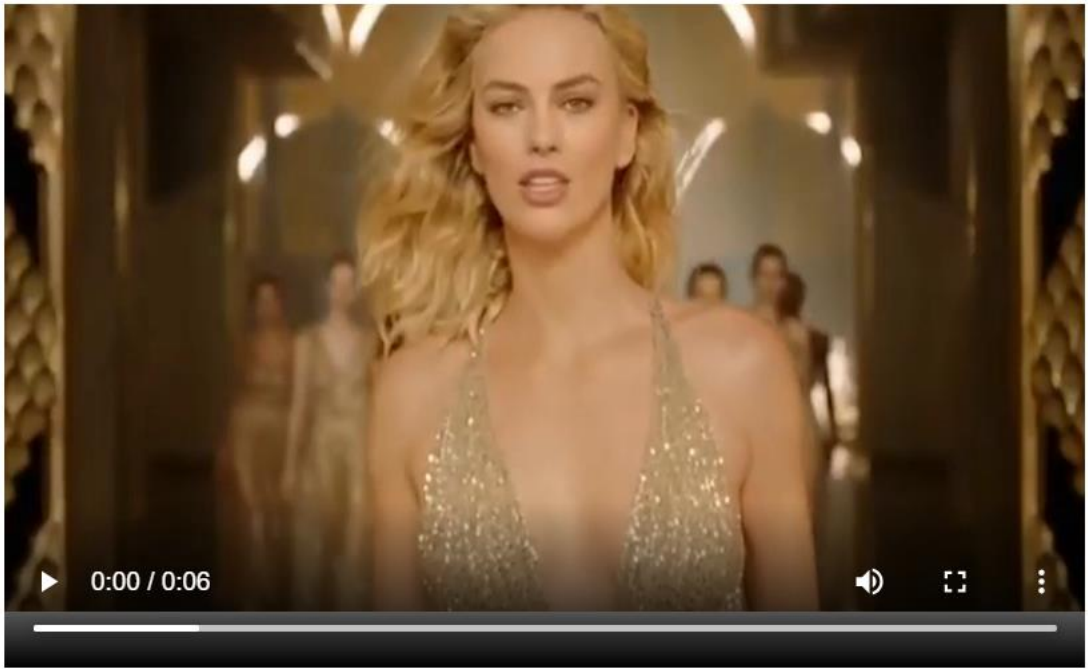
Zendaya (genuine). Speech: “And so it’s really interesting, because I can relate to that for sure, I mean I've been called controlling, I used to get in trouble... well I didn't really get it trouble because I was kind of a goodie two shoes in school”.

Stefan Löfven (genuine). Speech (translated): “And therefore, I will ask you once again for something very difficult, but very necessary. The little rest we had in summer and autumn is really over”.

Steve Carell (deepfake). No speech, but bells ringing.



Dua Lipa (deepfake). No speech, but saxophone playing.



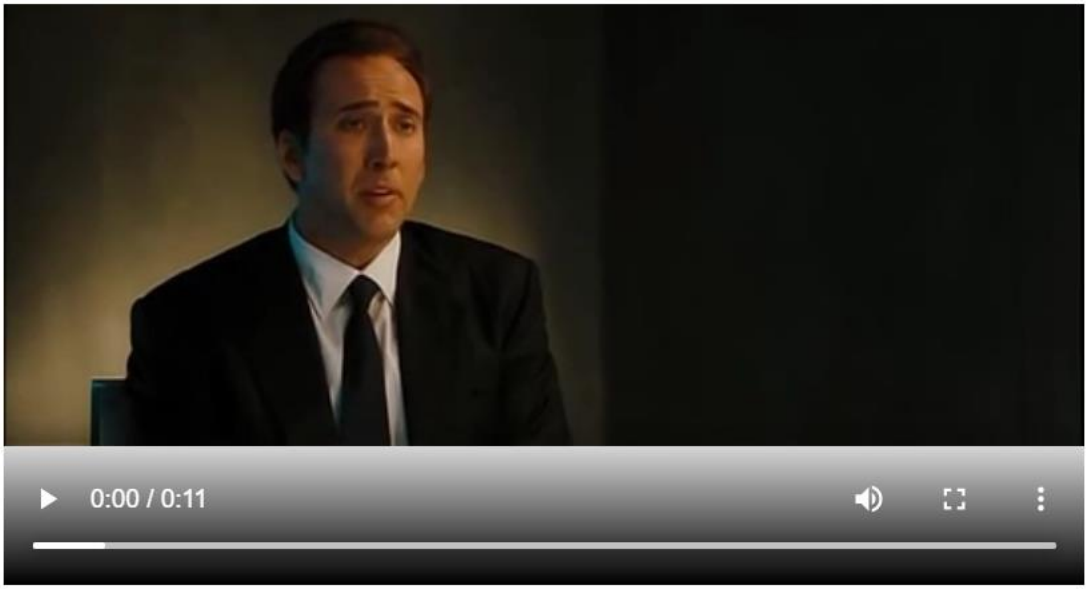
Margot Robbie (deepfake). Speech: “J’adore”

Jared Leto (genuine). Speech: "And just in that short amount of time, when I came out, there was a shutdown, a state of emergency and the whole world had changed".

Edward Snowden (genuine). Speech: "It can actually be a lot harder to remember a password that they tell you has to be 13 characters long, or something like that, has to have exclamation points, has to have numbers, has to have upper and lower case."

Kim Jung-Un (deepfake). Speech: "People are divided, your voting districts are manipulated, voting locations are closing so millions can't vote".

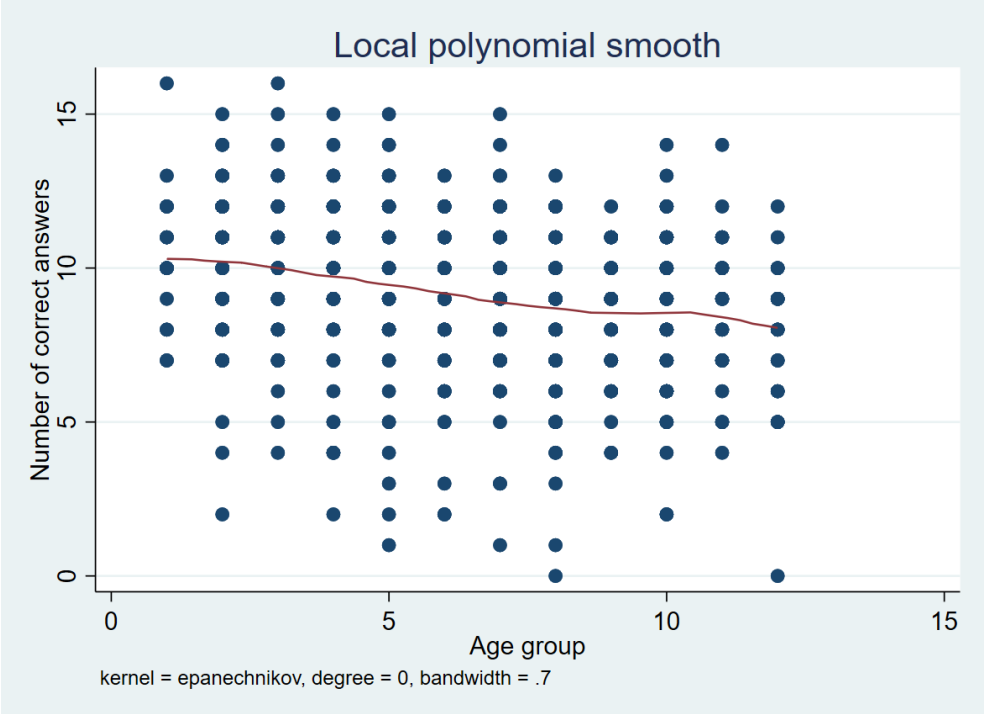
Melania Trump (genuine). Speech: "Strong, independent, very detail oriented, and staying true to herself".

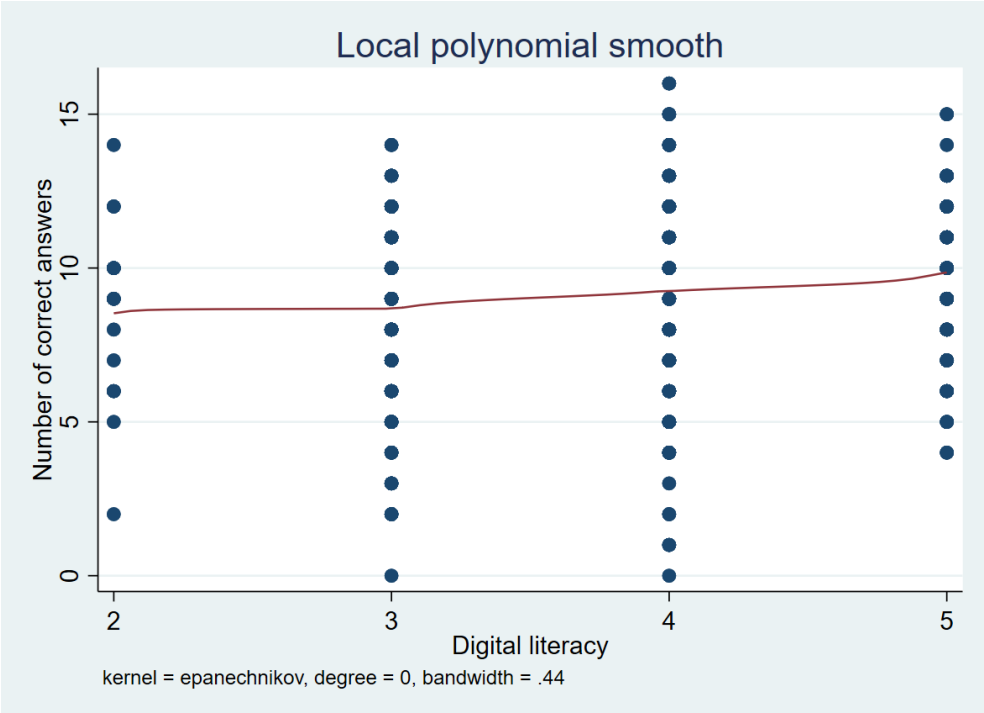


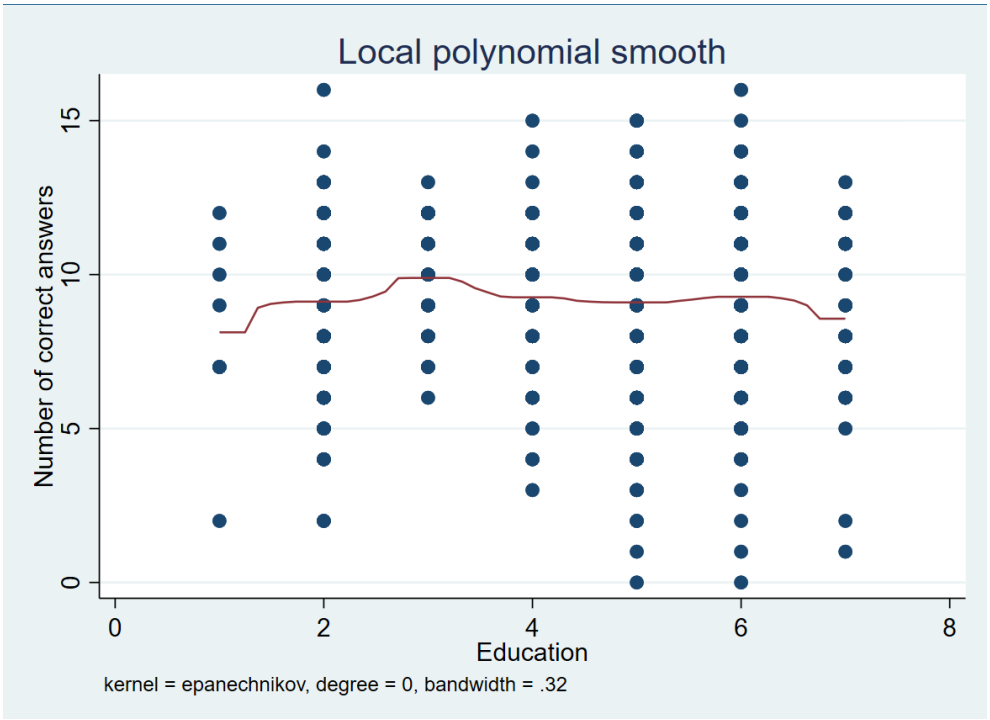
Nicolas Cage (genuine). Speech: “Tell me I'm everything you despise, that I am the personification of evil, that I'm, what, responsible for the breakdown of the fabric of society and world order”.

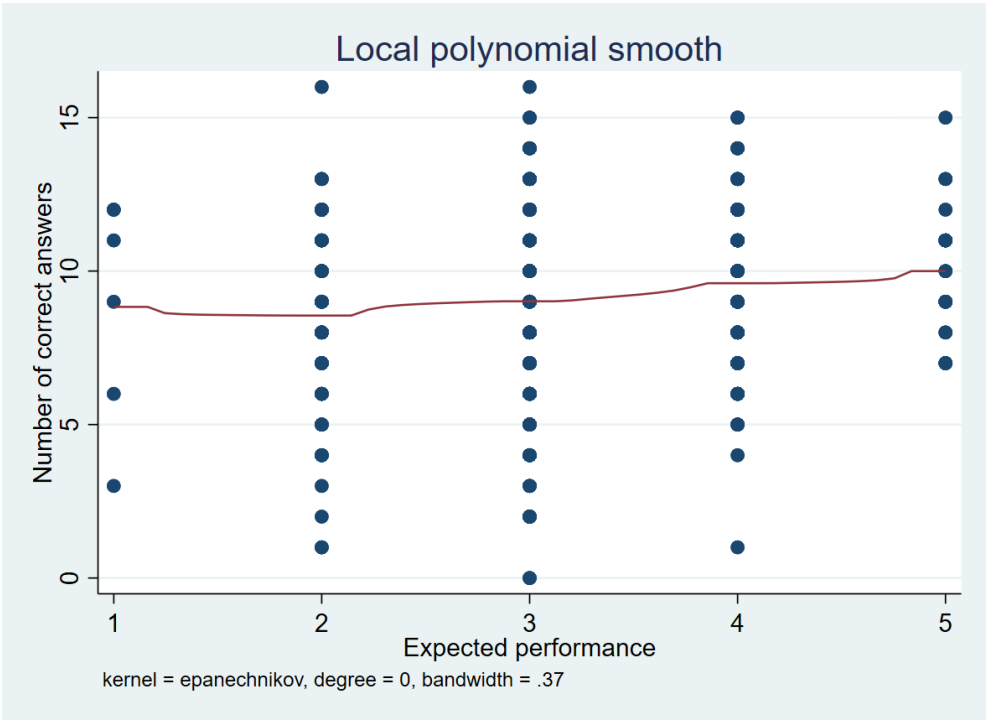
Mette Frederiksen (deepfake). Speech (translated): “The region of the capital is particularly affected. Here, the situation is now so serious that the authorities have raised the level of risk to level 5 in the warning system”.

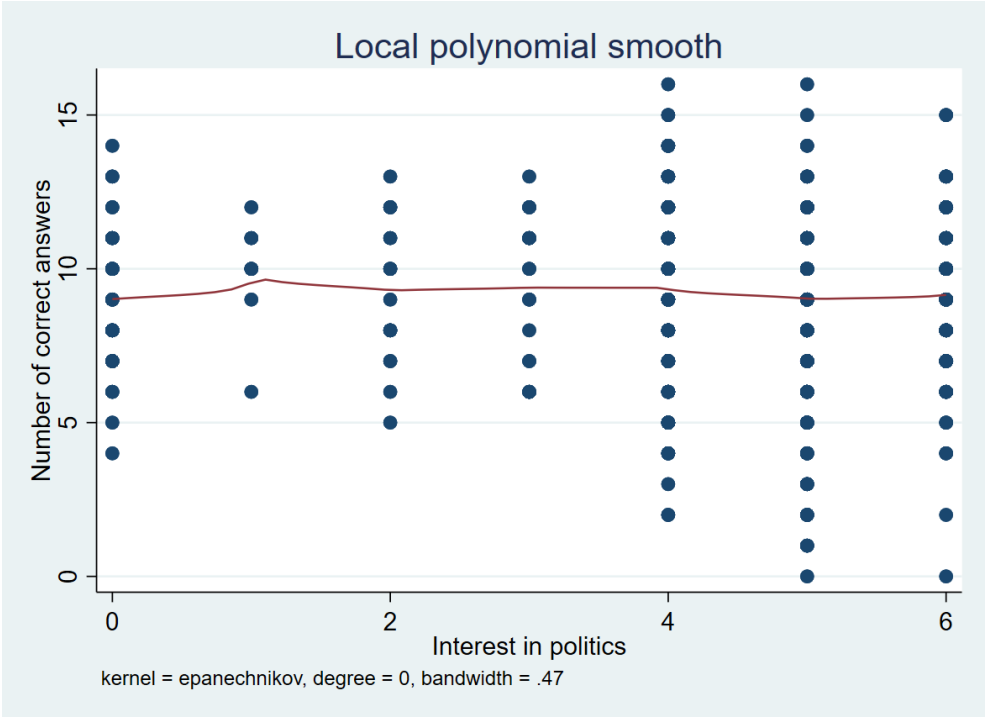
Attachment 3: Local polynomial regression for Classification of deepfakes

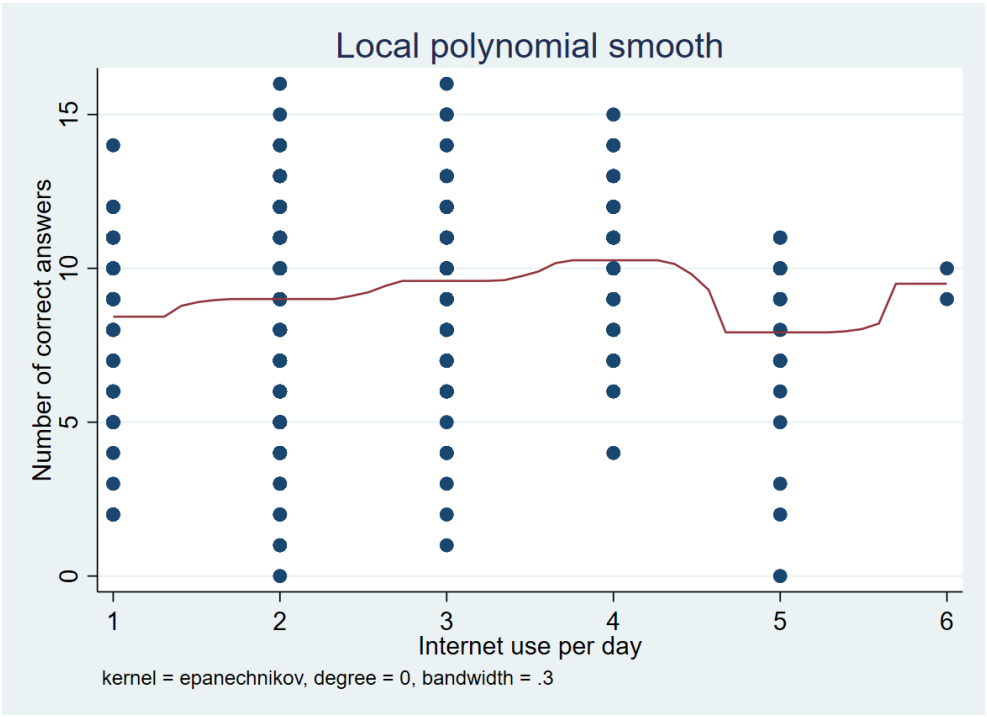


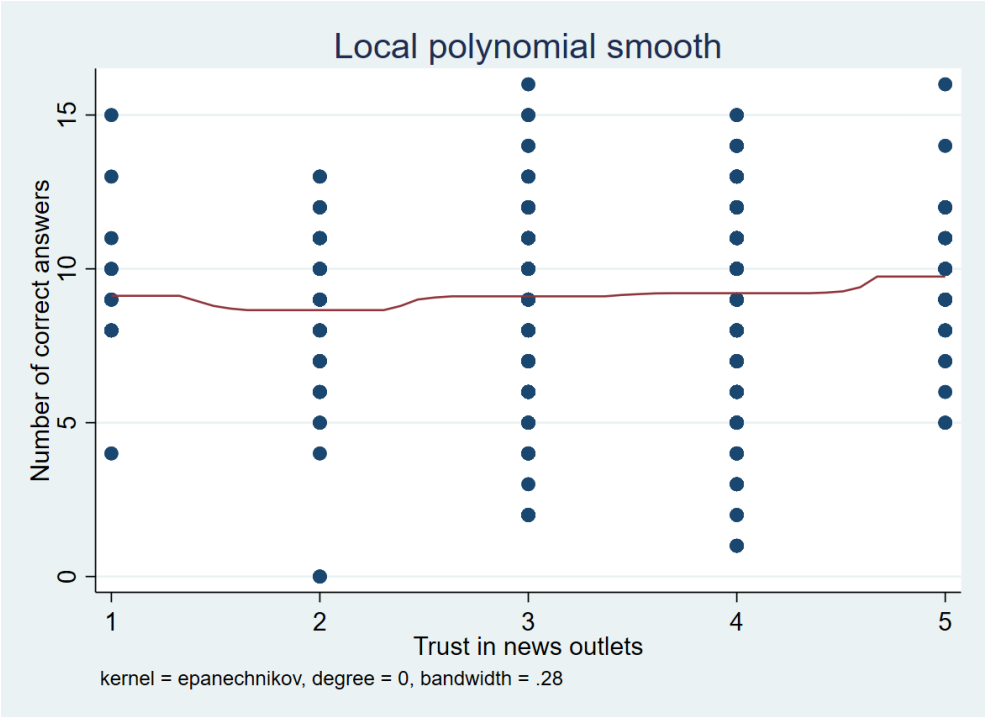



Local polynomial smooth
Number of correct answers
0
5
10
15
0
2
4
6
8
Education
kernel = epanechnikov, degree = 0, bandwidth = .32

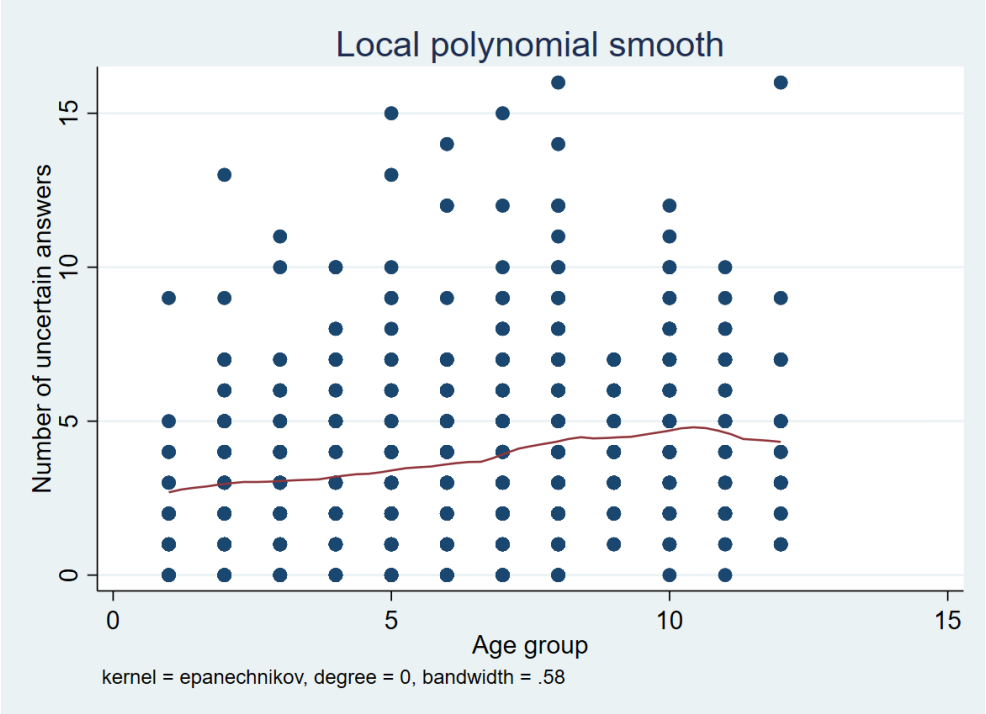

Local polynomial smooth
Number of correct answers
0
5
10
15
1
2
3
4
5
Expected performance
kernel = epanechnikov, degree = 0, bandwidth = .37

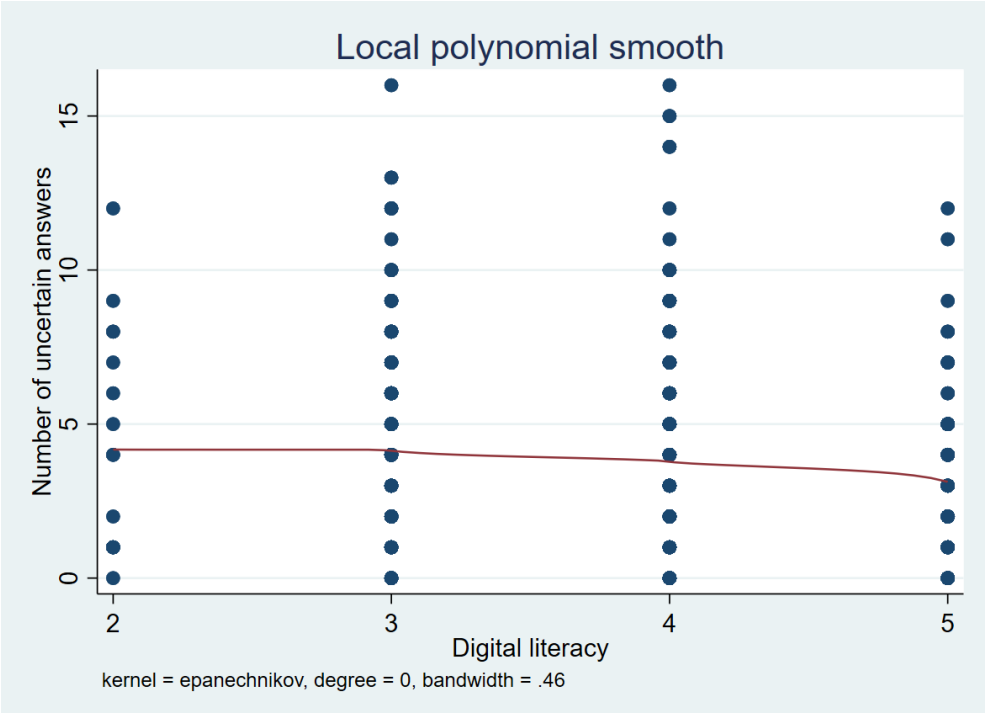

Local polynomial smooth
Number of correct answers
0
5
10
15
0
2
4
6
Interest in politics
kernel = epanechnikov, degree = 0, bandwidth = .47

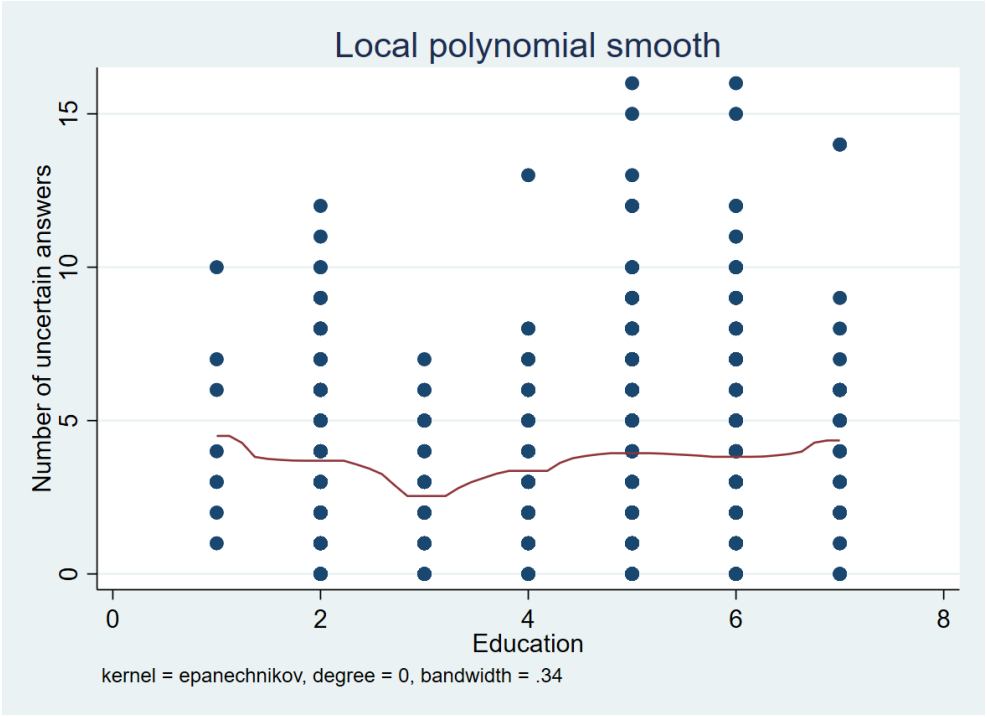

Local polynomial smooth
Number of correct answers
0
5
10
15
1
2
3
4
5
6
Internet use per day
kernel = epanechnikov, degree = 0, bandwidth = .3

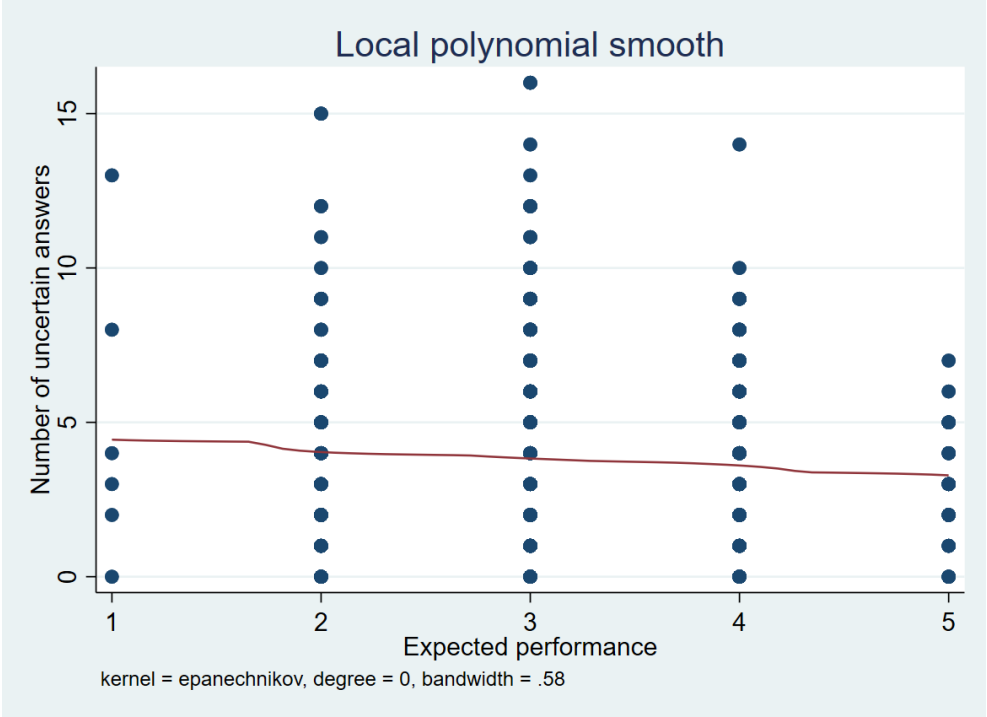

Local polynomial smooth
Number of correct answers
0
5
10
15
1
2
3
4
5
Trust in news outlets
kernel = epanechnikov, degree = 0, bandwidth = .28

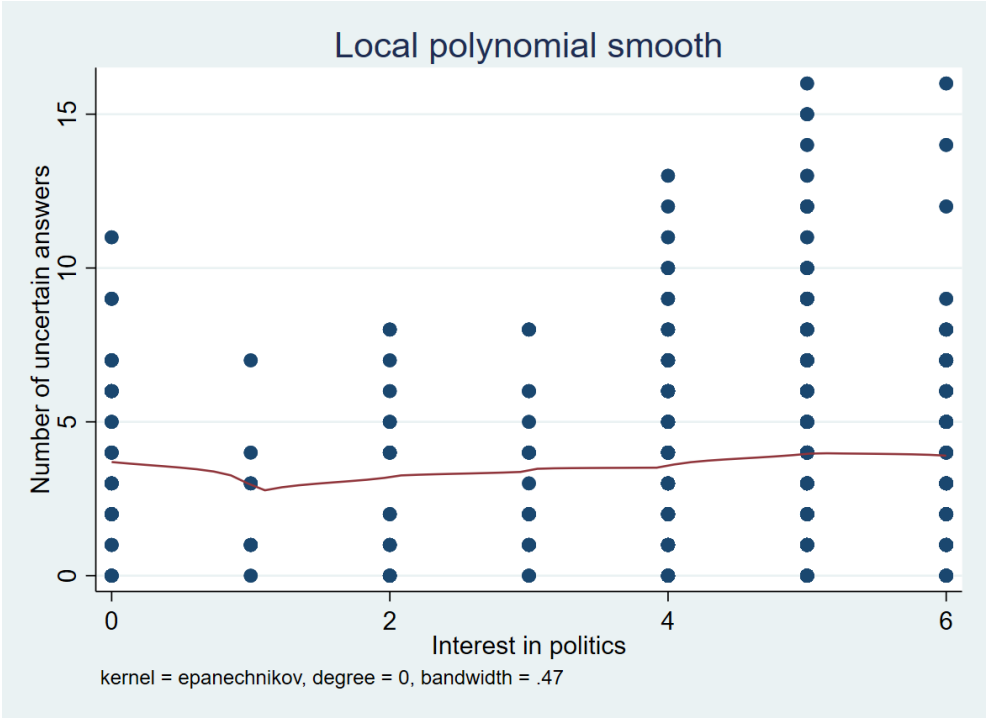
Attachment 4: Local polynomial regressions for Uncertainty when classifying

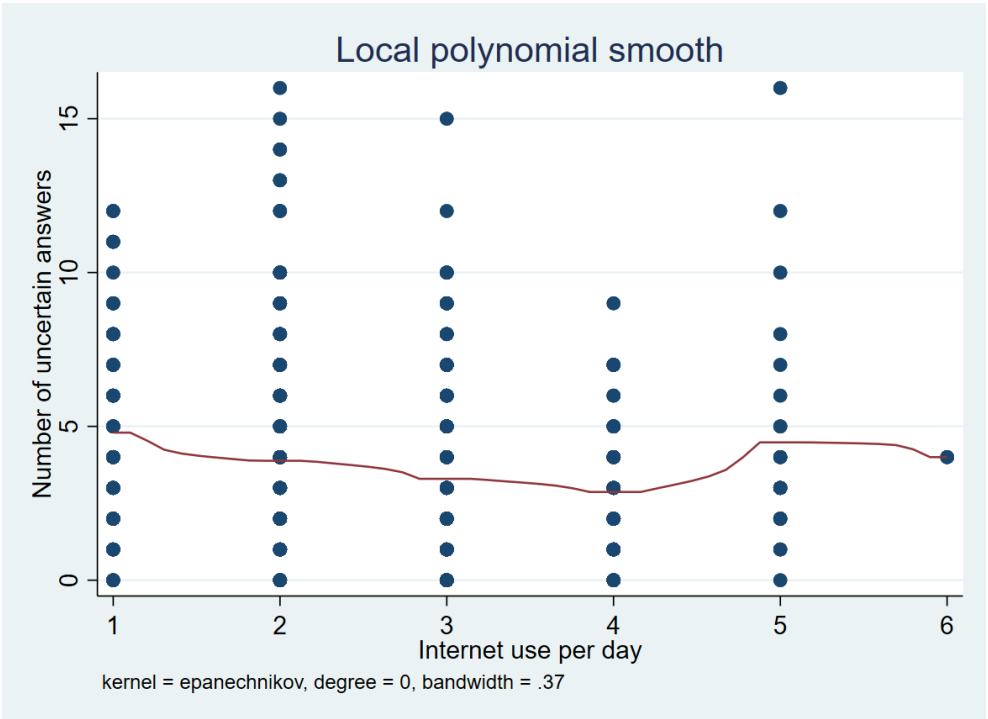


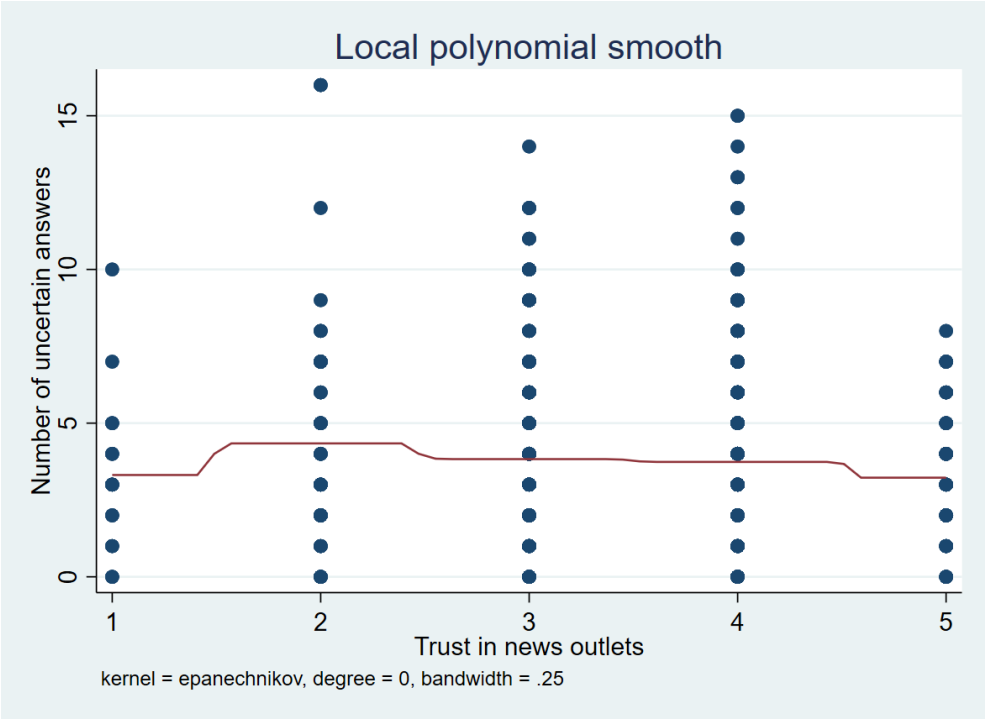



Local polynomial smooth
Number of uncertain answers
0
5
10
15
0
2
4
6
8
Education
kernel = epanechnikov, degree = 0, bandwidth = .34

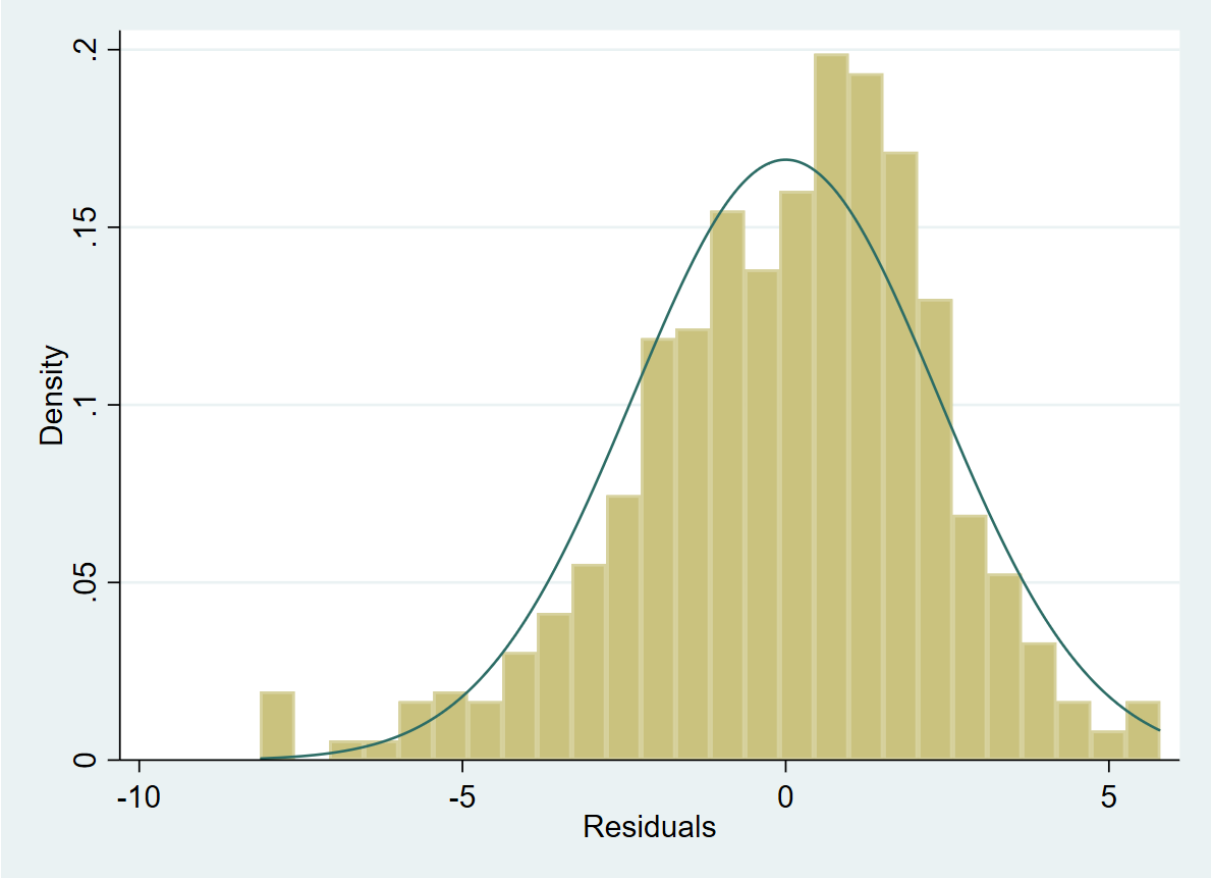

Local polynomial smooth
Number of uncertain answers
0
5
10
15
1
2
3
4
5
Expected performance
kernel = epanechnikov, degree = 0, bandwidth = .58

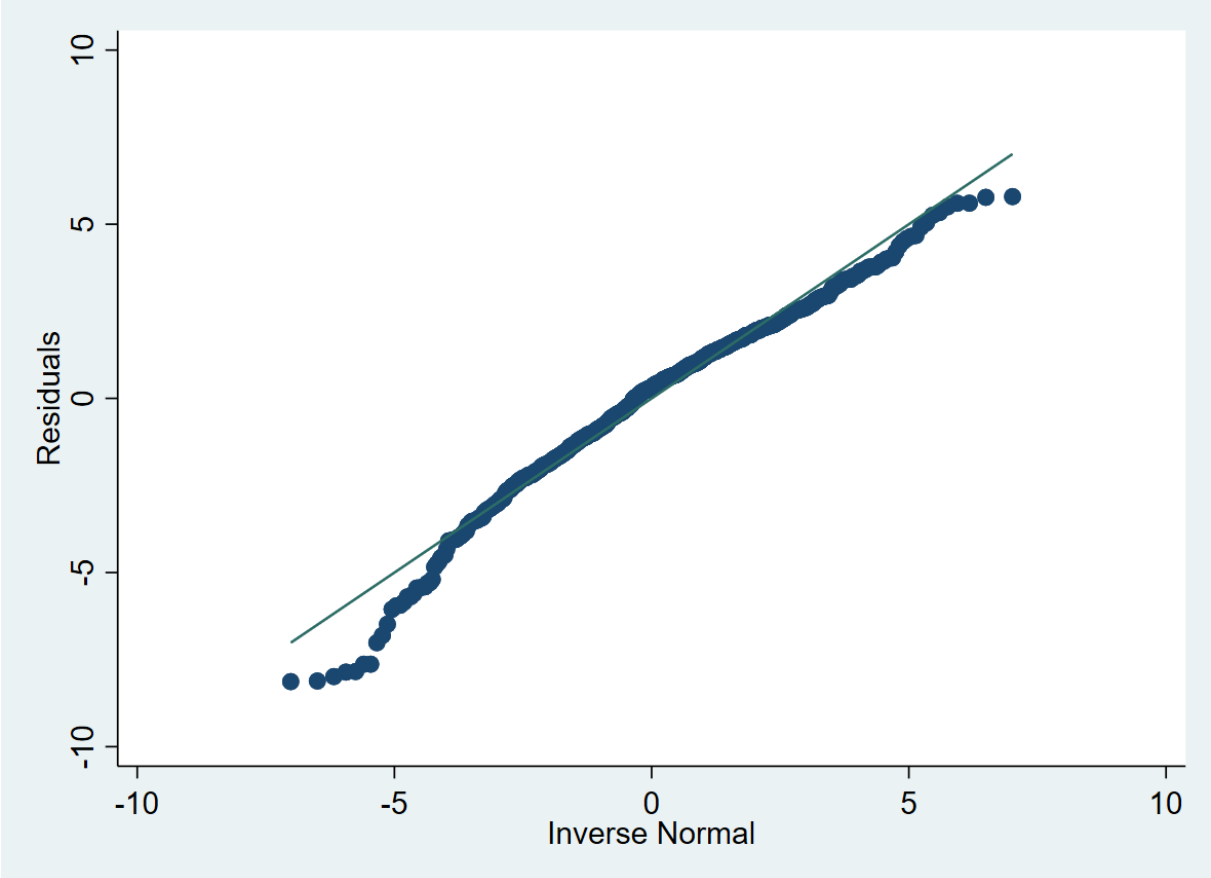

Local polynomial smooth
Number of uncertain answers
0
5
10
15
0
2
4
6
Interest in politics
kernel = epanechnikov, degree = 0, bandwidth = .47

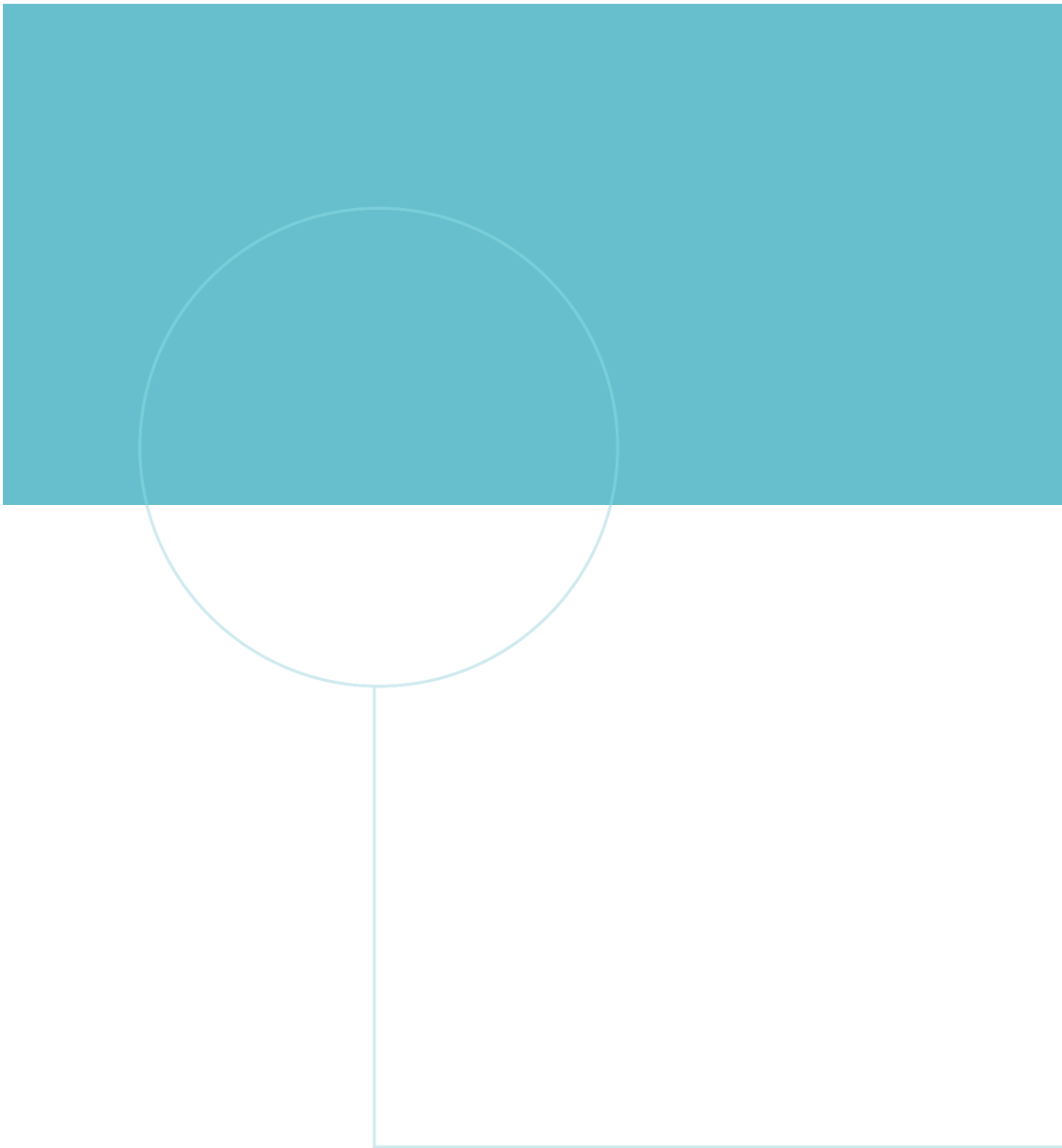

Local polynomial smooth
Number of uncertain answers
0
5
10
15
1
2
3
4
5
6
Internet use per day
kernel = epanechnikov, degree = 0, bandwidth = .37


Local polynomial smooth
Number of uncertain answers
0
5
10
15
1
2
3
4
5
Trust in news outlets
kernel = epanechnikov, degree = 0, bandwidth = .25

Attachment 5: Relatively normally distributed residuals







NTNU

Norwegian University of
Science and Technology