Uppsala university
Inst. for informatics and media

Social Media's Take on Deepfakes: Ethical Concerns in the Public Discourse

William Bogren & Mohamed Abdul Hussein

Course: Thesis
Level: C
Term: Autumn 2023
Date: 2024-01-19
Supervisor: Carl-Mikael Lönn

Abstract

The rapid advancement of artificial intelligence has led to the emergence of deepfake, digital media that has been manipulated to replace a person's likeness with another. This technology has seen significant improvements, becoming easier to use and producing results increasingly difficult to distinguish from reality. This development has raised ethical discussions surrounding its deceiving nature. Furthermore, deepfakes have had a considerable impact and application on social media, enabling their spread. Despite this, the public discourse on social media, along with its societal and personal values associated with deepfakes, remains underexplored. This study addresses this gap by examining social media discourse and perception surrounding the prominent ethical concerns of deepfakes, and situating these concerns within the broader landscape of AI ethics. Through a qualitative method resembling netnography, 320 posts from Reddit and Youtube were thematically analyzed through a passive observation, along with their respective comment section. The findings reveal various concerns, surrounding misinformation and consent to deeper fears about deepfakes' role in fostering distrust, as well as more abstract apprehensions regarding the technology's abuse and harmful applications. These concerns further revealed how generalized established AI ethical principles might be interpreted in the deepfake context, also showing how and why these principles might be violated by this technology. Particularly it revealed terms how principles such as dignity, transparency, privacy and non-maleficence might be diverged in deepfake applications.

Key words: Artificial Intelligence, Deepfake, Ethics, Social Media, Public discourse, Netnography

Table of Content

1. Introduction

This section introduces the topic of deepfakes and identifies the key issue this study addresses: the practical implications of deepfakes and the existing theoretical gap in their ethical understanding. It outlines the research objectives, questions, and the contribution and scope of this study.

1.1. Background

The digital age has seen a significant transformation in media and communication, in which digital information can easily be created, copied, communicated, and read globally (Chen, 2013). Moreover, today's digital media have also been influenced by advancements in artificial intelligence (AI). These AI technologies have further reshaped content creation, distribution, and consumption. For instance, present applications of AI within digital media can be found in chatbots, message optimization, and AI-generated content (Chan-Olmsted, 2019).

AIs rapid emergence is affecting society in many other ways and its technology is integrated into various aspects of human life, such as self-driving cars and for identifying criminals. With its constant advancements and impact, debates on the ethics around this technology and how it can be beneficial or potentially harmful have become more and more pressing (Liao, 2020). For instance, AI-driven facial recognition technology is widely used in security systems, enhancing safety but also raising privacy and surveillance concerns. The discussion around AI in digital media is no exception, while it may have enabled more personalized and engaging media experiences, it has also brought forth ethical questions (Floridi et al., 2018).

Within the ethical discussion, private companies, research institutions, and public sector organizations are actively developing principles and guidelines for ethical AI. These frameworks address not only the internal mechanics of AI but also its societal complications and the ethical use of these technologies. As such, guidelines and principles range from inherent biases in the AI algorithms to the usage responsibility of these systems not to harm society. Although there's a general consensus on the need for AI to be ethical and on the fundamental principles it should follow, the rapid development of these technologies still leads to an ongoing discussion regarding what constitutes ‘ethical AI’. This debate extends to the specific ethical obligations, technical standards, and best practices necessary for its implementation (Jobin et al., 2019).

Among the prevalence of AI and in the discussion of its ethics, one emerging area is “deepfakes” (Karnouskos, 2020). A deepfake is a digitally altered media that is generated by manipulating a person's likeness or replacing it with another through the implementation of AI techniques. There are multiple different types of deepfakes, such as face-swapping in videos and voice-swapping in audio, which are often combined (Kietzmann et al., 2020).

The term "deepfake " combines "deep learning" (an AI method) with "fake" and was popularized by an anonymous Reddit user who shared manipulated videos of celebrities in adult film clips (Kietzmann et al., 2020). Similar to its origin, a large majority of deepfakes online are still pornographic and have a strong connection with digital media, especially social media, where they find a vast audience. Given that text, images, videos, and sound are

the foundation of the interaction on these platforms, the impact that deepfakes can have in this arena is significant (Karnouskos, 2020; Maddocks, 2020).

Several manipulated videos have been circulating on popular social media platforms and gaining attention. An early example of deepfake technology is the 2018 video of Barack Obama, where he ironically addresses the risks of deepfakes (BuzzFeedVideo, 2018). Additionally, the rise of TikTok has provided deepfakes with a new platform to spread on. For instance, a deepfaked tiktok-video featuring Tom Cruise dancing has amassed a remarkable 93 million views to date (DeepTomCruise, 2022). Within solely audio deepfakes, platforms like NotJordanPeterson.com have also drawn attention by allowing users to generate speech in the voice of Jordan Peterson, a famous psychology professor and author, although the site was later shut down following legal issues (Cole, 2019).

However, social media platforms not only host a significant portion of this content but also serve as vital platforms for sharing viewpoints, discussion, debate, and the formation of public opinion and public discourse on different issues (Neubaum and Krämer, 2016; Kou et al., 2017), such as deepfakes. This is highlighted in the reactions to the Barack Obama video. Within the conversations in the Youtube videos comment section, there are remarks on how the technology can be used as a tool for framing (BuzzFeedVideo, 2018), highlighting the growing concern and awareness among the public.

Deepfakes carry both potential benefits and consequences. As highlighted by Chesney and Citron (2019), they might have educational purposes, artistic value, and foster autonomy and self expression. However, while deepfakes can be seen as creative, humorous, and educational, like in the case of the videos imitating Tom Cruise and Barack Obama, they often do so without the permission of the individuals whose images or voices are used and can be hurtful. Furthermore, deepfakes are inherently used to manipulate truth, leading to the possibility of spreading false information (Westerlund, 2019). As a result, academics have raised conversation surrounding the ethics of this technology, ranging from consent, misinformation, privacy, and reputational sabotages (Karnouskos, 2020; Westerlund, 2019). Likewise, regulators also have concerns about the implications of the technology, its societal value, and what desirability lies within its regulation (van der Sloot and Wagensveld, 2022).

Moreover, deepfake technologies have evolved to become more realistic, believable, and especially accessible, raising future concerns that anyone can fabricate deepfakes that are virtually indistinguishable from authentic media (Kietzmann et al., 2019). Existing deepfake source code and algorithms are already often open source for anyone to use. Furthermore, for those without coding expertise, computer apps, service portals, and marketplace services exist to aid in creating deepfakes (Ajder et al., 2019). For instance, there are online sites where anyone can create professional looking profile pictures on oneself through AI-manipulation (Kudhail, 2023).

1.2. Problem Discussion and Problem Description

Deepfakes, with their increasingly convincing and realistic nature, are becoming more difficult to distinguish from genuine media and easier to create. With these advancements and the ease of spread on social media, the ethical implications of deepfakes have become significant and pose a real-world problem. They range from potential harm from misinformation and the distortion of truth to violations of privacy and consent, especially when individual likenesses are used without permission.

As for the AI-landscape as a whole, there is an evident need for ongoing awareness and discussion regarding the ethics and disruptive use of AI-technologies. While some existing ethical principles are agreed upon, the ongoing debate around these topics shows that there is still disagreement on how to properly approach AI-ethics (Jobin et al., 2019). As the rapid development of these technologies is a factor in this uncertainty surrounding AI-ethics, it can also indicate that the growing evolution of deepfakes also contributes to this issue.

The current problems surrounding the field of AI-ethics is further remarked by scholars like Hagendorff (2019) and Munn (2023), who find these existing principles “meaningless” in the sense of being incoherent, omitted, contested, difficult and not applied in practice. Moreover, they believe that AI ethics and systems are still not developed in alignment with societal values. Hagendorff (2019) proposes that ethical approaches should move towards being situational rather than overgeneralized, and not only focus on technology but the social and personal aspects as well. This highlights the necessity to transition from general AI ethical standards to a more focused discussion on the ethics of specific AI technologies, such as deepfakes, within their applicable contexts, like social media.

While there are discussions and researches surrounding deepfake ethics, with considerations by Karnouskos (2020) and Kietzmann et al. (2019), they overlook the unique context provided by social media discourses. As further outlined in ”2.1.2. Deepfake Ethics and Existing Research”, the specific ways in which deepfakes are discussed and received on social media platforms seem to remain underexplored in existing research. This gap in research fails to capture the public perception and discourse surrounding deepfakes in the very arenas where they are most frequently encountered and spread. As emphasized by Karnouskos, social media has become an integral part of deepfake complex dynamics, which is expanded on in the section “2.1.3. Deepfakes in Social Media.".

Adopting Hagendorff’s (2019) perspective further underlines the importance of considering how users on social media perceive deepfakes when discussing the ethics of this technology. While discussions on these platforms may not fully address “ethical technology” aspects, such as potential biases or the transparency of the algorithms behind deepfakes (Jobin et al., 2019), they do offer valuable insights into the social and personal impacts of AI. These user perspectives can reveal societal and personal values, which are crucial aspects that Hagendorff identifies as missing in current approaches to AI systems.

To further clarify, these user-perspectives give a real-world context from the people that engage with deepfakes, revealing individual reactions and perceptions towards the technology and therefore giving attention to the personal implications of it. Users share their opinions and can express how they might be personally affected by deepfakes, offering valuable insight

into their experiences and concerns. Additionally, these discussions extend beyond personal viewpoints, contributing to the broader public discourse on deepfakes. The decentralized and interactive nature of various online platforms facilitates an environment where users can freely engage in open dialogue (Neubaum and Krämer, 2016). As further explained by Kou et al. (2017), the setting of these dialogues allows participants to discuss and evaluate different viewpoints and arguments. They continue discussions until they are satisfied that the best reasons have been fully expressed and supported. During these social media conversations and debates, the author highlights that participants aim to understand each other's viewpoints and are open to changing their initial opinions if they encounter stronger arguments. Consequently, this public discourse can both reflect and shape the societal response and its values regarding deepfakes.

In essence, the discourse on social media platforms can address a critical gap identified by Hagendorff (2019) in current AI ethics: incorporate societal and personal values into the ethical framework. Furthermore, these values and insights are not only just part of any public discourse, but embedded within the specific context of social media where deepfakes are not only present but actively integrated. It is in this arena where users engage with and respond to this technology, making their interactions and perspectives particularly significant.

Addressing this theoretical gap and deepfakes practical implications, the problem of this study lies in the need for a deeper understanding of social media reactions to deepfakes, their ethical considerations, and broader social and personal implications. By acknowledging how these AI-generated manipulations are discussed and perceived on social media, it can discover personal and societal values in this public response to the real-world challenges they present.

1.3. Research Question and Purpose

The purpose of this study is to explore and understand the discussions and public perceptions of deepfakes on social media platforms. It aims to gain an in-depth understanding of the prominent ethical concerns posed by deepfake technology, providing insights into significant societal and personal values. Furthermore, the study examines how these concerns are situated within the broader conversation about AI ethics. It specifically seeks to investigate how these concerns can be interpreted in light of established AI ethical principles and, simultaneously, how the notions of these principles are reflected and potentially be violated in the context of deepfakes.

<u>The research questions for this study are:</u>

Q1. *What prominent ethical concerns exist within social media discussions and perceptions regarding deepfakes?*

Q2. *How can the identified concerns be interpreted by established AI ethical principles, and indicate instances where these principles may not be adhered to?*

1.4. Knowledge Contribution and Knowledge Characterization

This study aims to deliver an in-depth understanding of social media perceptions and discourse regarding deepfakes. It focuses on inductively identifying and thoroughly analyzing the main concerns within these discussions as an important part of the societal response and attitude to this technology. This can reveal new ethical perspectives on the complex dynamics of deepfakes from users that encounter the technology. Particularly it can highlight the personal and societal values of AI usage, which are argued to be often overlooked in traditional AI ethical approaches.

In bridging the theoretical and practical aspects of AI ethics in the context of deepfakes, the study also contributes to the broader conversation about AI ethics. By analyzing concerns surfaced in social media discussions through the lens of established AI ethical principles, the research deepens the understanding of how these principles are applied, interpreted, and potentially challenged by deepfakes. It explores two main areas: firstly, how these principles can be contextually interpreted for deepfakes and secondly, assessing where deepfake technology can violate these ethical principles. This approach responds to criticisms that existing AI ethical frameworks are overly generalized, aiming to situate and apply these abstract principles within the specific context of deepfakes.

The anticipated findings of this research can provide valuable insights for further prescriptive research and actions, offering insights for policymakers, regulators, and social media platforms. Using this knowledge and understanding on social media and public perceptions, these stakeholders can develop more informed strategies and guidelines that resonate more effectively with societal concerns and expectations, contributing to the responsible development and use of AI technologies.

1.5. Scope

Regarding the scope of this study, it should first be emphasized that the "prominent" concerns will be explored. The main goal is not to cover all possible perspectives within the broad discussion around deepfakes on social media. Rather, it prioritizes a qualitative, rich, and in-depth exploration of those concerns that are most prominently and naturally present in the discussions and the meaning behind them. Therefore, it's important to note that certain less prevalent viewpoints may fall outside the scope of this analysis.

Lastly, given the ethical framework to interpret these concerns, this study adopts the ethical principles outlined by Jobin et al. (2019) overview of the AI-ethical landscape. While this approach provides a structured and comprehensive basis for analysis, it has limitations in the context of this study. This framework has a large focus on the technical aspects of AI, such as its development. This includes inherent biases as well as injustices in the algorithms themselves. In this study, the focus lies in the social and personal aspects these principles highlight in the use of AI, rather than the technological aspects of deep learning algorithms. Further, it should also be highlighted that other important ethical considerations such as media ethics, may provide important perspectives into societal values in the context of deepfakes but are out of the scope of the research's focus.

2. Theory

This section presents the theoretical foundation for this study, beginning with an exploration of the AI-ethical landscape. Established AI-ethical principles will be identified, that consequently will be applied to the analysis of the concerns found in the social media domain. The focus then shifts to the subject of deepfakes, examining their functionality and application, particularly within social media. Finally, the section reviews existing research on deepfake ethics, setting the stage for further investigation into this emerging area.

2.1. AI-Ethics

Artificial intelligence (AI) has rapidly evolved and seen a significant impact on our society in many ways. It increasingly influences more and more aspects of our society, from our personal lives to business and governmental tasks. While AI can improve the quality of life, it has also raised important questions about how it should be approached and what consequences its spread is bringing. In response to this, multiple scholars, institutions, and organizations have created initiatives to set ethical principles, guidelines, and frameworks for the application of these technologies to aid their development in line with improving society (Jobin et al., 2019). To guarantee the appropriate and ethical use of AI, it is important to investigate and comprehend the collection of ethical frameworks which have evolved throughout time. These frameworks provide useful insight on AI system development, implementation, and management. They include a broad variety of principles, such as justice, openness, responsibility, and privacy. In the next part, it will explore in further detail the way these frameworks provide valuable insights and lead to the ethical use of AI technology.

2.1.1. AI-Ethical Principles

As AI-ethics is getting more and more recognition, the ethical landscape widens. As presented in an overview of the ethical field of AI in the Jobin et al. (2019) study, they identified and analyzed 84 different frameworks from around the world. This extensive review revealed a diverse range of perspectives but also showed a consensus on several core principles crucial for ethical AI development and application. These principles can be summarized as follows:

Transparency: This principle highlights the significance of transparency in AI systems. It includes open communication on how AI systems operate, the data they utilize, and the reasoning behind their judgements. Transparency is critical for establishing trust and accountability, especially in institutions that have a large influence on people's lives.

Justice and Fairness: The concept of justice and fairness refers to how AI should help create a fair and just society. This concept emphasizes the need of designing and deploying AI systems in a way that does not reinforce current inequality in society, but rather promotes fair results.

Non-maleficence: This principle, which is based on the medical ethic of "do no harm," relates to AI by emphasizing that this type of technology should not hurt persons or society. It advocates for thorough assessment of possible negative consequences as well as the

deployment of safeguards to avoid damage. This concept is also linked to invasion of privacy, discrimination, physical damage, and breach of trust.

Responsibility and accountability: This principle holds AI systems accountable for their outcomes. It holds AI creators, programmers, and implementers responsible for the social, ethical, and legal consequences of their technology. It also involves responsibility to ensure that AI technologies are utilized in ways that respect human rights and dignity.

Privacy: Because AI systems often handle large volumes of personal data, privacy is a key challenge. This concept emphasizes the significance of safeguarding individual privacy and protecting personal data from unauthorized access and being exploited.

*Beneficenc*e: This concept involves maximizing AI's beneficial influence on people and society, such as improving human health, generating socioeconomic possibilities, and ensuring environmental sustainability.

Freedom and Autonomy: In the context of AI, this concept emphasizes the right to free speech, self-determination, and privacy. It promotes enabling people to make educated choices concerning the use of AI technology.

Trust: Developing trust in AI ethics requires building trust in AI technology and their inventors. It advocates for trustworthy, responsible AI systems and advises that education and transparent design be used to foster trust.

Sustainability: This principle advocates for environmentally responsible AI development, emphasizing AI's role in ecological preservation, social equity, and the creation of sustainable systems.

Dignity: Dignity in AI ethics relates to respecting and preserving human rights and individual value. It includes avoiding harm and dehumanizing practices while promoting respect for human dignity in AI development and application.

Solidarity: Solidarity in AI ethics concerns the implications of AI on social cohesion, particularly in the labor market. It underlines the importance of a strong social safety net and the equitable distribution of AI benefits.

Even with the general consensus on these aspects, the study also highlights that AI ethics is dynamic and constantly evolves in sequence with the rapid development of AI technologies themselves. Nevertheless, the principles identified by Jobin et al. (2019) paints a clear and comprehensive picture of the ethical principles of AI. In result, this generalized overview itself can be utilized in this study's reflection on how deepfakes situate in the AI ethical landscape. While these principles outline important aspects of AI and its algorithms' technology and development, there are several aspects to how AI should be used and applied to improve society. Within the analysis, the focus falls on precisely how these societal aspects correspond to the personal and societal values found in the concerns found on social media.

2.2. Deepfakes

As previously described, “deepfakes” are digitally altered media that are generated by manipulating a person's likeness or replacing it with someone else's (Langguth et al., 2021). This is done through the implementation of the AI technique of deep learning, a set of machine learning methods. These methods analyze large datasets to study and mimic a person. For instance, the end result can produce a video or image where one's face or original voice is replaced (Chadha et al., 2021).

Since deepfakes began, various types of deepfakes have surfaced that now include alterations in both videos, audios, and photos (Lewis et al., 2020). Audio deepfakes for instance, are generated through voice cloning. This is in order to imitate someone else's voice or through text-to-speech technology that effectively creates audio that mimics a person’s speech patterns, intonation, and accent (Kietzmann, 2017). These are achieved by training the deep learning models on extensive databases of an individual's speech, allowing for the generation of audio clips that closely resemble the original speaker’s voice (Khanjani et al., 2023).

However, video deepfakes and particularly face-swapping, are the most emerging type. Face-swapping is when one person's face is substituted with another (Nirkin et al., 2019). This process utilizes advanced face detection to accurately register and synchronize facial gestures, lip movements, and other behaviors of an individual (Kim et al., 2018). When deepfake first was recognized on Reddit, face-swapping was the original type of content created (Yu et al., 2021).

With the rising awareness of deepfakes, several deepfake apps have also surfaced. One such example is the face-swap apps “FaceSwap” and “FakeApp”, which gained significant popularity and recognition (Leibowicz et al., 2021; Guarnera et al., 2020). These apps also represent the continually growing ease of use and availability of deepfake generation, making it possible for anybody to create very convincing counterfeit pictures, sounds, and films (Nguyen et al 2021). Deepfake algorithms and source code are often open source, but even without coding expertise one can easily generate deepfakes by using these types of apps or other websites and services online (Ajder et al., 2019). Further, deepfakes have not only become more accessible but they are also seeing a rapid improvement in their quality with each one becoming more credible and realistic (Kietzmann, 2017).

With its advancement, deepfake technology has garnered widespread attention and is influencing the media production landscape significantly (Langguth et al., 2021). It is seeing multiple usages as well as misusages. For instance, it has allowed players in video games to give their faces to their avatars and businesses have acquired tools to create customized messages. However, it has also led to instances where it has been used for identity fraud and to spread fake news to influence political campaigns (Kietzmann et al., 2017; Karnouskos, 2020).

But the most well-known usage for deepfakes has been for pornographic content (Ajder et al., 2019). One such example is a deepfake application called “DeepNude” that surfaced in 2019 and was explicitly designed to produce fraudulent explicit images of persons without requiring the subject's permission (Leibowicz et al., 2021).

As a consequence of the spread and misuse of deepfakes there have been multiple conversations about ethical issues. In the case of DeepNude, the technology presented significant privacy concerns. People have also raised concerns about its influence on the public at large, such as its misleading use in the case of deceptive political campaigns (Leibowicz et al., 2021). The capabilities of deepfakes and their ethical implications are further outlined in the section on “Deepfake Ethics and Existing Research”.

2.2.1. Deepfakes Relation with Social Media

In the case of deepfakes, social media has become a free market for the distribution of AI-generated content. Deepfakes leverage the way information is communicated, read, and acted upon to spread rapidly and widely to large audiences. As such, social media has become an integral part of the complex dynamics of deepfakes. Since social media platforms also influence society. It is evident that deepfakes therefore have a large impact (Kietzmann et al. 2017; Chesney and Citron 2019; Karnouskos 2020).

The impact can especially be seen in the nature of social media's contribution to virtuality, this can be seen in the astonishing example of the viral Tom Cruise deepfake (DeepTomCruise, 2022). According to Westerlund (2019), the speed and reach of social media enable convincing deepfakes to quickly reach millions. Social media platforms, with their algorithms optimized for user engagement, tend to promote content that resonates with users, regardless of its truthfulness. This aspect of social media design makes it easier for deepfakes to go viral, and even more so when they contain sensational or controversial content.

The virality aspect also creates the vulnerability of one individual or deepfake to have a significant impact and affect millions (Karnouskos, 2020). Deepfakes’ hyper-realistic and deceptive nature combined with the reach and speed of social media has allowed convincing fake news to reach audiences globally and negatively impact our society. With over 100 million hours of content watched on the internet daily, the ease with which deepfakes can be shared and the difficulty in verifying their authenticity make social media an ideal environment for the rapid spread of fake content and information (Chadha et al., 2021; Westerlund, 2019).

Another important aspect of social media in relation to deepfakes is that its content fuels the data needed to train the AI and generate media. Platforms like Facebook, X (formerly named Twitter), and Instagram provide an enormous amount of data that AI can use to create text, images, and videos. Therefore, deepfake do not only use social media to spread content but it's accessible information that can be copied also facilitates the creation of the altered media. This has significant effects on media production as sources such as a publicly available voice sample from a speech and a photograph or video from social media can together lead to the creation of a realistic deepfake video (Karnouskos, 2020).

2.2.2. Deepfake Ethics and Existing Research

Deepfakes ethical considerations and positive and negative aspects have been extensively researched in the field and by multiple scholars. In this section, the existing research field of deepfake ethics is outlined while also highlighting important ethical aspects raised by these studies.

Firstly, it should be noted that many scholars have researched the many applications deepfakes can have. A literature review by Mahmud and Sharmin (2021), that explores the deepfake development process and current capabilities, also discusses the spectrum of applications. Relating to the positive aspects of deepfake implementation raised in this study, the author highlights that the technology's accessibility has led to innovative uses in art, film, and advertising. For example, the Dalí Museum used deepfake technology to create an interactive experience with Salvador Dalí, and in the film industry deepfakes have saved significant time and resources during the editing process. The authors also highlight that the technology has also been beneficial in other sectors like retail, news, and healthcare, offering personalized experiences, deepfaked news anchors, and improving the protection of patient data.

On the other hand, the multiple negative applications of deepfake technology is where ethical implications prevail. Deepfake has gained significant attention around its misuse against individuals, especially celebrities and political leaders (Mahmud and Sharmin, 2021; Kietzmann et al., 2020; Widder et al., 2022). As highlighted in a study from 2019 by Ajder et al., deepfakes are today most commonly used for pornographic purposes. According to their research, 96% of online deepfakes are non-consensual pornography, whereas 99% depict women celebrities. Further, the study also highlights that deepfakes have impacted other notable domains, primarily politics where fabricated videos of world leaders have surfaced.

Notable ethical implications have been highlighted in this type of application of deepfakes. Firstly, manipulated pornographic content has been seen as a breach of privacy and consent. Secondly, the political sphere shows concerns on the nature of deepfakes to manipulate information (Karnouskos, 2020). Manipulation of truth and the consequences of deception such as misinformation, disinformation, and fake news are central topics within deepfake ethics that are raised by multiple scholars, such as the literature review by Mustak et al. (2023) and document analysis by Karnouskos (2020). Within this aspect, Karnouskos also examines how it has effects on crimes such as fraud and wider social consequences such as trust in journalism.

The ethical implications show consequences from deepfakes that affect multiple actors. This is especially outlined in a study by Diakopoulos et al. (2019) on ethical deepfakes in the context of elections. In their research, they developed eight hypothetical scenarios to assess the potential impact of deepfakes on the 2020 U.S. presidential elections, using a participatory approach involving crowd workers. In their analysis, they categorized the potential harm deepfakes can inflict on viewers /listeners (recipients of the deepfakes), subjects (the targeted persons whose likeness are used for the deepfakes) and social institutions, which the authors describe as *"the domain in which a deepfake operates"*. In line with this, Mahmud and Sharmin (2021) emphasizes that deepfakes can have harmful

consequences for the whole society, both in the short and long term. Interestingly, they also highlight that the people at risk are the ones regularly using social media.

Kietzmann et al., (2020) also discuss the ethical implications of deepfakes from different perspectives. Their paper explores the possibilities, danger and consequences of deepfakes and proposes approaches to mitigate their risk. However, their study divides the positive and negative into individual, governmental, and organizational aspects. One important aspect raised by Kietzmann et al. is that not only celebrities have become subjects and sustained substantial damage from deepfakes. Within the authors' individual perspective, one example was raised about a young, non-famous woman who found herself among a huge amount of adult content where she was face-swapped onto the bodies of porn actresses. This demonstrated how anyone might be impacted by nonconsensual use of adult content or online harassment which can damage one's reputation, emotional welfare, and career. The author also discusses how misuse of the technology can lead to various forms of abuse, including revenge porn, identity theft, and bullying.

Considering the other two categories of Kietzmann et al. (2020) study, they highlight that organizations also can suffer damage to their reputation from manipulated content. In the governmental perspective, the consequences are as before focused on the political sphere. Here, the authors discuss how deepfakes might lead to the fabrication of evidence, manipulation of public opinion, and interference in democratic processes.

The earlier studies discussed the application, capabilities, and implications of deepfakes. However, Widder et al. (2022) investigate the social standpoints from the perspective of the creators that spread open-source deepfake tools and source code. As a result of their study, the authors found that these creators and participants do not perceive ethical responsibility as lying with themselves or the technology. Rather, the ethics of deepfakes solely depends on the user and how they utilize it.

To summarize this, scholars take different approaches to research in the context of deepfake ethics, where they highlight several ethical implications. From the quantitative research proposed by Ajder et al. (2019) to the literature and document review by Karnouskos (2020) and Mustak et al. (2023). Existing research also discusses and categorizes different perspectives on these implications. While discussing the ethical implications these can have for individuals and whole societies, there is an absence of empirical studies that actually investigate how society and individuals perceive or value this technology. While studies like Widder et al. (2022) to some extent capture this in their study on the creators standpoint, the public discourse and especially in social media is not fully captured.

3. Research Design

The following section will be devoted to the description and motivation for the research strategy and methodology chosen for this study. Further, limitations, choices, and ethical considerations necessary to conduct the study will be outlined.

3.1. Research Methodology

To capture the social media discourse surrounding deepfakes, this study adopts a qualitative research methodology that is significantly influenced by the principles of *netnography*. This research method, initially coined by Kozinets (2010), is used for exploring online communities and cultures. Netnography is a digital adaptation of ethnography to an online environment, enabling the study of social interactions and behaviors within digital communication. Traditionally, ethnography involves immersive engagement in a social environment with regular observations to understand participants' behaviors within that context. Researchers listen, participate in conversations, and conduct interviews to gain a comprehensive understanding of the group's culture and behaviors (Bryman, 2018). Netnography extends these principles to the internet, adapting them to the unique dynamics of digital interactions. Within this digital domain, data can be collected from occurring public conversations, including both textual and multimedia communications in their contemporary channels (Kozinets, 2010).

Since the purpose of this study lies in distinguishing discourses on social media, it is evident that the most effective approach is to directly investigate these online spaces and the communication within them. Aligned with the summarization by Berg (2015), netnography allows for the use of the internet as an arena to conduct research on the internet.

Using the aspects of netnography in this study allows for the collection and examination of data from various social media platforms. In these environments, users discuss, share, and perceive deepfaked content or engage in topics regarding deepfakes. Therefore, this approach allows the study to thoroughly capture valuable viewpoints and interactions in these digital spaces, which can give insights into the ethical implications that users see in this emerging technology.

3.2. Implementation

Netnography encompasses a variety of approaches to consider when implementing. Like ethnography, it's an umbrella methodology that can integrate multiple methods such as observations, interviews, and textual analysis (Kozinetz, 2010). This study adopts a non-participatory observational method. In this approach, researchers only observe and interpret online communications without directly interacting with or influencing the community (Bertilsson, 2014). As highlighted by Costello et al. (2017) this method also offers rapid and cost-effective data collection compared to more traditional qualitative research methods like focus groups and personal interviews.

However, the selection of this passive approach is not only grounded in convenience. It is in line with the goal of analyzing existing and natural social media discussions, rather than creating them. As highlighted by Bertilsson (2014), this method reduces the risk of researcher bias and ensures the authenticity of the social interactions observed. This passive presence is necessary to capture unobtrusive information and an unbiased view of the public discourse where users naturally engage with and respond to deepfake content within their social media environments.

However, it should be noted that this non-participatory stance motivates the labeling of the approach as "significantly influenced" by netnography, rather than fully embracing it. Costello et al. (2017) point out that the traditionally recognized requirement for active human involvement in netnographic research is gradually making room for non-participatory methods. Even so, the authors note that certain studies have similarly refrained from defining their work as netnographic due to a lower level of researcher engagement with participants than what's considered essential for authentic netnographic studies. The consideration behind this approach can still be seen in line with a later study by Kozinets (2015), where he emphasizes that "you may have perfectly good reasons for lurking in the shadows".

Nevertheless, the strategy still remains significantly influenced by the principles and framework of netnography and its online research. The passive stance, while diverging from the traditional requirement for active researcher engagement, still embodies key aspects such as the analysis of digital communications and the exploration of online communities. This study therefore draws extensively from netnographic methodology while adapting it to suit the specific research needs and constraints.

In regard to the research process, this study still follows the fundamental netnographic phases as described by Kozinets (2010). However, as Costello et al. (2017) have pointed out, certain aspects of the traditional netnographic methodology, specifically the activities of establishing a cultural entry and obtaining member feedback, are not as applicable in the context of a passive research approach. Figure 1 illustrates the steps for the study process, which are adapted from Kozinets (2010). In the subsequent sections, these steps will be discussed as part of the research strategy.

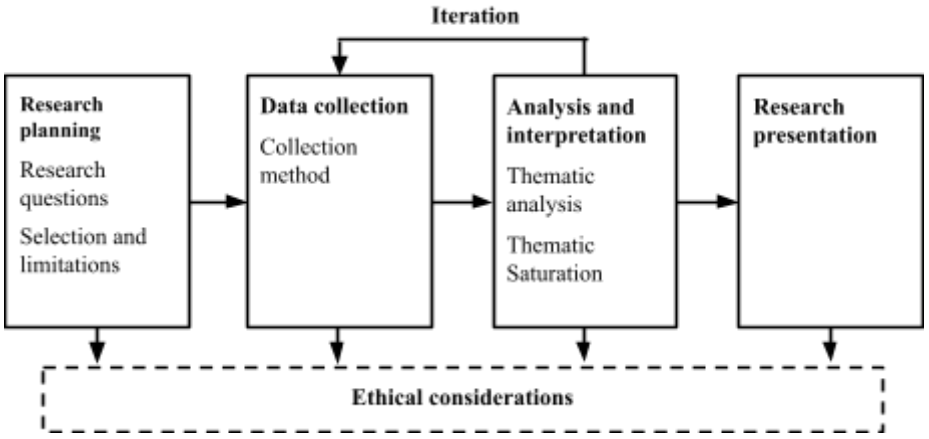


Figure 1. Research Strategy Process. Phases Adapted from Kozinets et al. (2010) Netnographic Flow.

3.2.1. Research Planning

The initial phase of the netnographic study focuses on research planning. This step involves defining the research field and formulating specific questions aligned with the study's objectives. This step is essential for narrowing the focus and guiding the data collection efforts efficiently (Kozinets, 2010). By establishing research questions, limitations, and selection relevant to the investigation, it sets a clear and targeted framework for the study. This phase not only guides the investigation within the social media domain but also ensures that the data collection is purposeful and aligned with the goals of the research.

Research Questions

The following research questions guide the investigation into the ethical implications and public perceptions of deepfakes on social media platforms. The questions are designed to explore various aspects of deepfake technology and contexts, dividing the research objective within ethical considerations, societal impacts, and user perceptions for an in-depth understanding.

1. What is the social media user's awareness of deepfakes?
2. How do social media users perceive deepfakes?
3. What ethical concerns are predominantly raised in social media conversations about deepfakes?
4. Which context and domains are concerns and viewpoints surrounding deepfakes made (for instance, entertainment or pornography)?
5. What is the consensus around these concerns? Are there any contradictions around the viewpoints?
6. What are the implied consequences of these concerns, and who are these concerns affecting?

Selection and Limitation

Berg (2015) lists three boundaries to be established when defining the empirical field in a netnographic study: temporal, relational, and spatial. The temporal boundaries are concerned with the duration and timing of the study. The relational boundaries pertain to the interactions and relationships between participants in the study and the researchers. Lastly, the spatial boundaries encompass questions of who, where, and what will be studied.

In terms of the temporal boundary, the data collection and analysis will be limited to two weeks (29th November to 13th December). However, as later motivated, the data being investigated and analyzed during this period will consist of data from the past year (archival data).

Regarding relational boundaries, the passive observation approach minimizes the risks of ongoing relationship-building with subjects during the study. Further, the data subjects will be fully anonymous, both for the researchers and in the presentation of the results, ensuring that no personal information is gathered or recognized.

Finally, considering spatial constraints, it becomes crucial to strategically choose relevant platforms and sources for an effective analysis of discussions across the extensive social media landscape. For this study, the process involves a multi-layered approach. First, suitable

social media platforms need to be identified. Secondly, specific posts on these platforms need to be pinpointed. Thirdly, a particular focus should be placed on the comments and replies within the posts. Comments and replies are vital as they involve the main discussion to be examined for the thorough understanding of social media discourse and how users perceive deepfakes. However, posts also add to the discourse and set the foundation for the comments, and the dynamics and demography of the platform also affect the content and conversations. Therefore, criterias and assessments for all these layers (platforms, posts and comments and replies) need to be systematically determined.

When selecting online communities or sources, Kozinets (2015) highlights multiple aspects that should be considered. He emphasizes that the ideal online sources for investigation should be relevant to the research focus and questions, active with recent and regular communications, interactive with significant communication flows among participants, substantial in terms of having a critical mass of communicators, heterogeneous with a variety or similarity of participants, and data-rich, offering detailed or descriptive content. However, Kozinets also states that “it can make good sense to trade off one or more of these criteria”.

Drawing from these considerations, the selection and its criterias for the spatial boundary is visualized in Figure 2 and further described below.

Data sources	**Criterias**	**Selection**
Social Media Platforms	- *Diversity of platforms and content*	Reddit, YouTube
↳ **Posts**	- *Social Media Metrics* - *Existing Discussions* - *Relevant to research question* - *Mass of communicators, dara richness*	Uploaded in the past year Minimum of 100 comments 80 post (iterative)
↳ **Comments & Replies**	- *Relevant to research question* - *Social Media Metrics* - *Resonation with users* - *Representative opinions and concerns*	Minimum of 5 like

Figure 2. Selection of Data Sources (Spatial limitation)

Social Media Platforms

Firstly, diversity of platforms is considered. In this study, the social media platforms will consist of Reddit and Youtube for their varied perspectives and content on deepfakes. Youtube offers direct insights into public perceptions of trending deepfake-related videos, such as the Barack Obama deepfake, while Reddit provides rich and diverse textual discussions on the topics of deepfakes. These platforms enable different behaviors and discourses, therefore providing varied perspectives on deepfakes. Further, they have a large user base, ranging from over 1.1 billion on Reddit as of 2022 (Dixon, 2023a) to approximately 2.7 billion on YouTube as of 2024 (Shepherd, 2024), allowing a critical mass of communicators to be involved in the discussions.

Further regarding demographics, YouTube's user base consists of 54.4% male and 45.6% female, with the largest demographic being Indian with 462 million users, followed by the U.S. with 239 million users. YouTube's reach is extensive across various age groups, with a

significant presence in the 19-32 age range, yet evenly distributed among internet users (Shepherd, 2024). Reddit, on the other hand, consists of 62% male and 38% female users (Dixon, 2023b). As of April 2023, Reddit's audience is predominantly from the U.S., making up around 48% of its user base (Dixon, 2023c). Within this US audience, the average Reddit user falls within the 19-29 year age bracket (Shewale, 2023). While overrepresented in certain aspects, these demographics still highlight that a diverse range of users engage in these platforms. Nevertheless, these demographics are still both vital for understanding the context and perspectives within which deepfakes are discussed and how the societal and personal values in the findings might correlate to certain user groups.

Posts

Criteria for selecting posts include relevance to the stated research questions and social media metrics like the number of comments. Posts uploaded in the past year with a minimum of 100 comments are targeted. This limitation is based on the indication of a significant number of communicators and communication flow within the post. Additionally, it will allow for more varied perspectives from different participants. Further, the one year time aspect is needed to reach recent discourse while also allowing the acquisition of enough data. This means that the data collection is archival, where data that already exists is gathered. Since the research approach is passive, there are no limitations to real-time discourse.

Considering the amount of posts to gather, as highlighted by Medéia and Carlos (2019), determining a sample size is both vital and challenging due to the vast volume of available data in online research. Due to the multi-layered approach in the broad social media landscape, setting a predefined sample size that is representative is complex. The collected data needs to be manageable and still comprehensive in identifying prominent concerns. As a result, this study opts for an iterative approach. This procedure is further explained in “3.2.4. Iteration and Thematic Saturation”, but in short it means that 40 posts per platform will be collected until a satisfactory sample size is reached. While awareness lies in the fact that all viewpoints still might not be identified, iterating with 80 relevant posts per data batch ensures that prominent ones are identified.

Lastly, this study uses a simple random sample to mitigate any further biases, ensuring a fair representation of posts that meet the set criteria (Taherdoost, 2016).

Comments and Replies

The study will be focused on comments relevant to the research questions, resonating with users and representing a variety of opinions. Comments with at least 5 likes are selected to ensure they reflect shared viewpoints and prominent concerns. The threshold of likes serves as a benchmark for identifying comments that resonate with a wider audience. As highlighted by Neubaum and Krämer (2016), comments that receive higher engagement and likes indicate viewpoints or opinions that hold significance for other users, suggesting these comments are more likely to reflect prevailing attitudes or notable perspectives within the discussion. To further understand how these comments are reflected, the replies to these comments are also collected. However, replies to replies can not be gathered from the Reddit API and will therefore be out of scope for this study. While this implicates the comprehensiveness and depth of these discourses, this limitation simultaneously streamline

the focus of the study. Targeting the analysis of the resonating and prominent viewpoints and its direct interactions without going through a forest of replies and back and forth debates.

These purposive aspects of this multi-layered selection also have limitations. There are inherited risks of selection bias, which could lead to a skewed representation of public and social media discourse by prioritizing more sensational viewpoints and neglecting less prominent online spaces. This encompasses a range of elements that might be overlooked, such as social media platforms with different demographic profiles, comments and opinions with fewer likes that are perhaps controversial, and posts and topics that are less popular or mainstream. While this selection is necessary due to the constraints of the study, the main goal is after all not an exhaustive exploration of every perspective but rather a focused aim at an in-depth understanding of the main concerns in deepfakes.

3.2.2. Data Collection

As we initiate the data collection phase with the defined selection and limitations criteria, finding and acquiring relevant data revolves around filtering and investigating through the vast amounts of online content to identify content that is in line with the research objectives (Kozinets, 2010). Using existing APIs from Reddit and Youtube, the random sample of 40 posts from each platform will initially be gathered. As stated by Morsello (2017), the API protocol is one of the most efficient ways to collect data but is associated with legal and ethical issues that have to be considered.

The acquisition of posts will be based on the searchword “deepfake”, to capture the vast amounts of topics on the subject, and filtered to only those that have been posted within the last year. For Reddit, only posts that exist in public subreddits (subsidiary threads) will be collected. While this might complicate further selection bias, as stated in the section “3.2.5 Research Ethical Considerations”, it is essential to respect the privacy of these online communities. Further, each post will be manually evaluated to ensure its relevance and alignment with the research questions, particularly focusing on whether it stimulates a discussion around deepfakes. Any post that does not meet these criteria will be excluded and replaced with another until a total of 80 suitable posts are identified. Additionally, for each selected post, comments that have received more than 5 likes, along with their replies, will be collected. Lastly, the collected data is going to be translated into English, if not already.

In support of the data collection and the later outlined analysis, a specialized low-level interface application was developed, detailed in”'Appendix A”. Regarding the data collection, this tool is specifically tailored to streamline the retrieval process, aligning with the research strategy and criteria. It efficiently fetches posts from both YouTube and Reddit, including their associated comments and replies. Its functionality not only ensures systematic selection and potential replacement of posts but also aids in managing the substantial data set. This application is vital to maintaining the integrity and organization of the research data, facilitating a more focused and effective analysis.

3.2.3. Data Analysis

In order to conduct a meaningful analysis of the collected data, this study adopts thematic analysis as outlined by Braun and Clarke (2006). As described by the authors, this approach is based on the identification, analysis, and interpretation of patterns or themes within qualitative data. Through a thematic approach, the complex ethical dimensions of social media discourses surrounding deepfakes can be thoroughly examined. By finding and interpreting patterns within these discussions and perceptions, prominent ethical concerns can be identified and categorized into coherent and meaningful themes. The suitability of thematic analysis for the material collected is also reinforced by its ability to handle large volumes of qualitative data, which is a characteristic of social media platforms (Braun and Clarke, 2006; Kozinets, 2010). This method enables the study to effectively organize and interpret the data to draw meaningful conclusions.

Other qualitative data analysis methods could be considered, like sentiment analysis to examine the emotional tones in discussions about deepfakes, or comparative analysis to investigate how conversations differ across social media platforms. For example, comparing the information-focused discussions and elaborate narratives on Reddit with the visual storytelling and perception found on platforms like YouTube. Nonetheless, considering the aim of identifying the different concerns and their meanings surrounding deepfakes, thematic analysis offers an in-depth understanding that extends beyond examining emotional undertones or platform characteristics.

This also motivates a more latent approach to thematic analysis, not only looking at data at face value but also investigating underlying ideas, assumptions, and conceptualizations (Braun and Clarke, 2006). Given the nature of social media and the complications of deepfakes, there are uncertainties about whether the viewpoints of dialogues are explicitly articulated. As stated by Benamara et al. (2018), who examined social media language in the context of Natural Language Processing (NLP), social media possess specific language characteristics such as irony and slang, as well as non-linguistic contextual information like emojis that shape discourse. The authors also emphasize that texts in social media form part of a cohesive whole, with each text contributing to the overall meaning of the discourse, resonating with the multi-layered approach of this research. Therefore, a surface-level analysis of discussions surrounding deepfakes might not only fall short of capturing the meaning of the discourse but could also potentially be misleading. By employing a latent approach, the analysis aims to uncover the indirect, underlying attitudes and beliefs, providing a more comprehensive and nuanced understanding of the deepfake phenomenon. This includes remaining aware of the surrounding context of the observed content, such as current or previous events that might influence the discussions, as well as understanding the unique dynamics specific to each social media platform.

Further, this analysis opts for an inductive approach, meaning that it involves deriving meaning and identifying themes from the data without any preconceptions or specific theoretical interests (Braun and Clarke, 2006). This is in line with the exploratory nature of our study, as it does not require a predefined theoretical framework, allowing themes to emerge organically from the data. While the outcomes of this analysis are intended to be compared with existing principles of AI ethics, it is crucial that the theme's development

remains independent. This ensures an impartial interpretation of the social media discourse and the academic literature.

As seen in the process for the analysis, this study follows the six-step process outlined by Braun and Clarke (2006) for conducting thematic analysis, which includes:

1. Familiarizing yourself with the data by repeatedly reading it.
2. Development of initial codes, in which the analyst highlights interesting features of the data, with codes representing the most basic and meaningful segments of the data that can be analyzed in relation to the studied phenomenon.
3. Searching for broader themes by collating data and potential codes into thematic clusters.
4. Refining the themes through a review process, where themes may be combined or removed if there is insufficient data to support them.
5. Defining and naming themes to establish their identity within the context of the data.
6. Presentation of the themes that have emerged from the analysis.

In the data collection and analysis artefact detailed in "Appendix A", functionalities have been integrated to streamline thematic analysis. The application visualizes posts along with their comments and replies, allowing for efficient note-taking and coding directly on the content. This aids the initial two stages of thematic analysis. Additionally, the tool facilitates the creation and editing of themes, and the collation of codes within these themes, supporting the crucial steps of data collation and theme refinement in the analysis process.

After defining, naming, and presenting the themes, the final step involves comparing them to previous studies outlined in "2.2.2 Deepfake Ethics and Existing Research" and interpreting them through the lens of the AI ethical principles outlined in "2.1.1. AI-ethical Principles". Firstly, this will show how the themes provide new and unique insights into the societal and personal values from the findings, positioning the result of the study. Secondly, the connection between the central concerns identified in the study and the existing ethical principles, can assess not only their alignment but also how these principles are applied and followed in practice.

3.2.4. Iteration and Thematic Saturation

The data collection and analysis process in this study will be iterative, whereas the study will continue its sampling of 80 posts and analysis until saturation. Reaching saturation, as outlined by Hennink and Kaiser (2022) and also highlighted within the field of netnography by Kozinets (2010), is indicated by the point at which no new insights emerge from additional data and the information begins to repeat. Consequently, it suggests that the collected data or sample size is sufficient for a thorough understanding of the research subject. Hennink and Kaiser (2022) emphasize that in qualitative research, the concept of saturation is not just about collecting a large quantity of data but rather about ensuring the comprehensiveness of the data in relation to the research topic. For this analysis, this specifically relates to inductive thematic saturation, which Saunders et al. (2018) describe as the emergence of new codes or themes.

There are different perspectives on assessing when saturation is met. Hennink and Kaiser (2022) describe a stopping criterion approach, where for instance, the 'X' number of further analysis without new codes defines saturation. However, Saunders et al. (2018) argue that saturation does not occur, but researchers themselves should consider when there is enough data and whether sufficient depth of understanding has been achieved. Drawing from this, no predetermined stopping criterion or iteration count will be defined. To facilitate and clarify the decision for saturation, counts for new codes and themes in each iteration will still be outlined in the presentation of the result.

In terms of the analysis process, this iterative approach also suggests that new codes from later iterations can be applied to existing themes, without the need to collate the current data sample. However, within each data batch, it is important to refine existing themes and their identities with the new inputs. How the iterations are applied in the data collection and analysis in practice is further visualized in Figure 3.

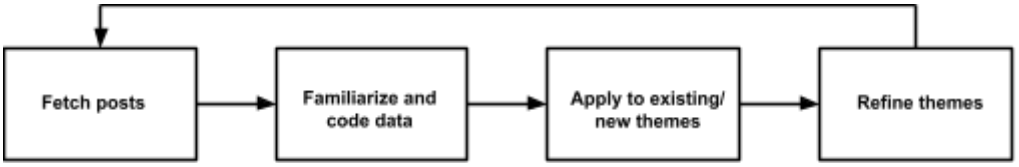


Figure 3. Iteration process

This approach ensures that prominent themes and their underlying factors are not overlooked and are thoroughly investigated. Although it may not encompass the entire range of discourses or codes to deepfakes, it can uncover subtler but equally important discussions that might not be immediately apparent. This contributes both to the depth and richness of the study's findings and the central concerns, while also remaining inclusive of other possibly meaningful viewpoints.

3.2.5. Research Ethical Considerations

In conducting this study on deepfakes in social media through a netnographic approach, several considerations need to be assessed regarding the ethical implications inherent in the research strategy. As emphasized by Kozinets (2010), ethical considerations in netnography are complex and not so straight-forward, particularly in the aspects of informed consent, data privacy, and the potential harm to study participants. Additionally, the ethical risks need attention in all the steps in netnography, from the selection of sources to the representation of the results. While general ethical guidelines exist, online research also needs to be assessed case-by-case, as advised by the Association of Internet Researchers (Franzke et al., 2020).

A primary concern in this study is the issue of informed consent. Given the vast amount of archived posts collected, informing and gaining consent is a difficult task. Franzke et al. (2020) notes that the use of APIs is specifically a problematic process since data is collected automatically and often in large volumes. As evidenced by previous studies in the same field, Proferes et al. (2021) conducted research on the ethical practices in studies involving Reddit, discovering that most did not seek consent in their data collection. Nevertheless, as Kozinets

(2015) points out, ethical considerations and the need for consent also need further assessment starting with what constitutes public and private data online.

The line between public and private is not so easily conceptualized, and many different perspectives exist on how to make a distinction. Unlike traditional netnography, which typically falls into human subject research and private exchanges (Kozinets, 2015), this study approaches the data as published text, which can be exempt from informed consent requirements. This perspective, suggested by Kozinets, is motivated in that data is archived, acquired from public spaces, and not subject to interaction and participation from the researcher. In line with this, this study strictly uses data from the public domains of social media platforms and avoids any private or restricted online spaces to respect users who might have an expectation of privacy. Further, it diverges from traditional netnographic approaches that might involve active participation or interaction in real-time discussions and instead focuses on passively collecting archived and pre-existing data. This approach and conditions, as Kozinets (2010) suggests, situate the collection and thematic analysis outside the conventional netnographic ethical framework that necessitates consent.

However, an obligation when treating online data as text rather than social interactions is to anonymize the identities of communicators (Kozinets, 2015). In this study, the personal identifiers are automatically deleted when storing data from the API, adhering to Franzke et al. (2020) guidelines. In line with Kozinets (2010) emphasis, this also includes direct quotes in the presentation of the result, which can easily be traced to users by full text searches in public search engines. In line with this, Proferes et al., (2021) exploration of studies on Reddit showed that almost 70% refrained from using quotes and the authors propose that researchers should acknowledge risk when using direct quotes where content is sensitive. While discourses around opinions and perceptions of deepfakes might not be viewed as sensitive, this constraint also supports the idea that the focus is strictly on the content and themes of the discussions rather than on individual users. Moreover, this anonymization can mitigate any discomfort for the users, who might not expect their contributions to be part of academic research, especially when no consent is acquired (Kozinets, 2010). Following this, it should be noted that the dataset used in this study will, for the same reason, not be shared.

Furthermore, this study is mindful of the potential harm and sensitivity associated with the interpretation of online discussions. The analysis aims to accurately represent the views and opinions expressed by social media users, maintaining contextual integrity and avoiding any misrepresentation or decontextualization of the data. This is crucial to uphold the integrity of the participants' contributions and to avoid sensationalizing or misinterpreting the discourse.

In addition to these netnographic-specific ethics, Franzke et al. (2020) highlights that the terms of service for internet spaces should be followed, and Kozinets (2015) suggests that the researchers national ethical institutions and laws should be observed. Considering the former, awareness is made for both Reddit and Youtube user agreements and terms of service for their respective APIs. As for the latter, the Swedish Research Council (SRC) is the authority assigned the position of informing research ethical questions (Regeringskansliet, 2015). SRC (2017; 2002) guidelines for non-participatory observation are in large parts in line with Kozinets (2015; 2010) and earlier remarks. For instance, they sympathize with Proferes et al., (2021) expressed exemptions in the need to seek consent or inform participants when using passive or “hidden” approaches, such as if the inconvenience is assessed not to be significant.

4. Results

In this section, the study results and analysis are presented. An introduction to the results of the analysis is first outlined, followed by the description of the identified themes.

4.1. Presentation of Results

The data collection and analysis resulted in four iterations and an examination of 320 posts, evenly split between Reddit and Youtube. The first iteration revealed the majority of underlining codes, 40 in total. Subsequent iterations contributed gradually with fewer additional codes, with the second iteration adding seven, the third two, and the fourth two. While the latest iteration added new codes, no new themes were discovered, and existing ones had no need to be refined. Drawing from this, the analysis and thematic exploration were assessed as saturated.

As for the content analyzed, it should be noted that a significant portion of the posts, discourses and concerns were centered around deepfake in the pornography context. This is consistent with the broader trend where pornographic content constitutes a significant majority of deepfake material (Maddocks, 2020). Likewise, the pornographic domain itself is also regularly discussed around its ethics (Verdier, 2018; Boynton, 2023).

Further, the analysis also revealed more specific cases surrounding deepfakes, which was significantly discussed and influenced the social media discourse on the technology. One of these involved the American Twitch streamer Brandon Ewing, also known as Atrioc. In January 2023, during a live stream, he accidentally showed a website selling deepfakes of other streamer friends (Cole, 2023). This “incident” was a popular topic within the Reddit space. Another notable case that was discussed primarily on Youtube was a recent instance involving the Indian Bollywood actress Rashmika Mandanna, where a deepfake video with her face morphed into another woman's Instagram video was widely circulated in November 2023. This incident led to a public outcry and a police investigation (Sebastian, 2023; Dixit, 2023). These two cases were widely posted and commented on in multiple online communities. Reflecting the demographics of these platforms, the case involving the Indian actress aligns with YouTube's predominantly Indian user base (Shepherd, 2024), while the incident with the American streamer corresponds to Reddit's largely American audience (Dixon, 2023c).

Looking past the case-specifics, four themes were identified from the concerns found in social media perception and discourses on deepfakes, as seen in the overview in Figure 4. Each is presented in subsequent sections.

Consent and Autonomy	This theme addresses the ethical concerns of using deepfakes to manipulate personal images without consent or awareness, highlighting issues of violated autonomy and the right to control one's own image, and the dehumanizing nature of non-consensual use.
Misinformation	This theme addresses the concern of deepfakes in spreading misinformation and manipulating truth, emphasizing their potential to harm and deceive individuals through the creation of false narratives and defamatory content.
Trust	This theme focuses on the broader societal impact and concern for deepfakes in eroding trust, leading to challenges in discerning reality and the potential shift towards a "post-truth" society where digital information is routinely questioned.
Abuse and Harmful Use	This theme addresses the concern for deepfake technology to be misused, and contrasting it to its beneficial application.

Figure 4. Themes for concerns in social media discourses and perceptions of deepfakes

Before going into the themes, some “honorable mentions” should be addressed. The analyzed discourses also heavily surround regulation and in some cases, accountability. The discussion surrounding regulation concerned on what laws and rules should be implemented in the deepfake context, and while connected to ethics, is not an ethical concern itself. The ethical concern in regulation can instead be found in accountability, however, only a few discussions were identified within this aspect. As a result, these two aspects are not considered in the presented result, even if they constitute valuable insights.

4.1.1. Consent and Autonomy

In the collected data, consent and autonomy were repeatedly discussed and emerged as prominent themes. These discourses concerned how deepfakes may be used for manipulation and exploitation of personal images and likeness without consent or the awareness of the subject, which also challenged the notion of individual autonomy.

In the perception of deepfakes, users expressed their discomfort, fear, and disgust at scenarios where individuals are depicted in AI-generated content without their approval. These comments reflected a violation of autonomy and consent and a deep sense of dehumanization, emphasizing the right of individuals to control their own image and body.

These concerns were especially highlighted within deepfake pornographic content, and particularly involving women. For instance, in the notable context of Atrioc, the streamer and the creator of the deepfaked content was widely condemned by users for violating personal boundaries. Further, since the deepfakes were bought from the platform, the profit motive

behind such exploitation further compounded the ethical violation. One important perspective was that the concern about consent was not only highlighted for the creator of the deepfakes but also for the consumers. In the case of Atrioc, he watched and did not create the deepfakes. Yet, many uttered contempt that he did not have consent to watch these videos of his streamer friends. In the discussion around consent in the pornographic context such as this, some users even compared the deepfakes to molestation or rape due to their non-consensual nature.

Conversely, some users questioned and contradicted whether a deepfake can really be seen as someone's body or their 'real' self and if it really is a breach of autonomy and requires consent. The main argument here was that deepfakes are just 'fakes' and imitations or representations. While there was generally a common discomfort with pornographic content, where the line should be drawn for consent was still subject to debate, both for adult content and in general. For instance, users questioned if deepfakes are disallowed, should they also ban satire, photo edits or painted pictures of someone?

Nonetheless, the discussions on social media often touched on the dehumanizing aspect of deepfakes. They were seen as reducing a person to an object for others' fantasies, stripping away their humanity. A part of this discussion was particularly raised around pornographic deepfakes involving children.

Lastly, while discussions on consent and autonomy were closely associated with pornographic deepfake content, noteworthy concerns also emerged in relation to creative professions, such as actors and music artists. For instance, there was distress around how a singer's voice could be stolen, and how it infringes on personal and intellectual property rights. Further, concerns were raised about the consent in using deceased actors' likenesses in deepfakes. However, whether and how the consent of an actor who has passed away can be obtained or inferred where expressed to be a big ethical question itself.

4.1.2. Misinformation

The thematic analysis of online discourses also highlighted several key concerns and insights into how deepfakes can manipulate the truth and deceive people, as well as its consequences. A primary concern for this lies in the convincing nature of deepfakes. Multiple users revealed that they themselves did not even realize that they were watching deepfake, even if it was expressed in the title. The comment sections repeatedly expressed alarm at the technology's ability to mimic reality so convincingly that it became difficult to distinguish truth from fiction.

This realisticness and the ease with which someone can create deepfakes created unrest in the discourse about how misinformation, propaganda, and fake news can be spread. Meaning that deepfakes enable distortion of information for people to gain from or harm someone. For instance, several users discussed how it will become easier to blackmail, frame, and trick people. This potential for misuse was especially concerning in political contexts, where deepfakes were raised as a potential tool to manipulate elections and distort democratic processes.

Several perspectives on who might fall to the consequences of misinformation from this technology were also expressed. For instance, deepfakes can make someone a victim by manipulating a video of you doing something you have not done, or as a viewer of the content, it can trick you into believing it. Further, the discussion highlighted how social media, with its ease of spreading information, can amplify its effect and reach more people, especially vulnerable groups.

These vulnerable groups were highlighted as older generations and children, who might be less familiar with technology. These people were seen to tend to more often believe in what they see, making them more susceptible and potentially targeted to being misled by deepfakes.For instance, a couple of posts discussed alleged scam ads that used the likeness of Mr. Beast, a well-known YouTuber, in deepfakes. The discourse showed concerns around targeting misinformation and exploitation of deepfakes since the Youtuber's target group was seen as younger children.

Further, misinformation and deception especially regarded reputational harm. Many comments pointed out that anyone can fall victim to deepfakes and how they might be used to harm someone's image by twisting their truth. One instance brought up is the case of revenge porn and the real-world consequences this might have, such as affecting one's job opportunities and personal life. On the contrary, deepfakes were also highlighted in some aspects as a tool for deception to benefit one's reputation, such as the use of "de-aging" deepfakes to appear younger, which can for instance be used for catfishing.

While many stated how convincing the deepfaked content was, a significant number of users also commented that they were able to easily identify flaws. Further, claims were made about existing software that efficiently detects deepfakes. Yet, there are still concerns about deepfakes constantly improving and if a point will come during the AI-detection and AI-generation race where it gets too realistic even for these tools to keep up.

4.1.3. Trust

The consequences of deepfakes also showed how they might reshape our understanding of trust in digital content and society as a whole. The realisticness and possibility of misinformation raised bigger concerns about trust in digital communication, media, and their content. Discussions signified how deepfakes will foster doubts about whether or not digital videos, images, and audio are authentic. Claims were made that distrust will continue to grow as these manipulated videos become more prevalent, potentially leading to a shift where digital evidence is routinely questioned. There were concerns that the manipulation of videos through deepfakes might become normalized and lead to a "post-truth" society where digital evidence is routinely questioned, creating significant challenges in discerning reality.

Another aspect of trust in social media is not the content consumed but the content created and shared. Multiple users made remarks about stopping posting and even removing their multimedia on social media platforms in fear of being deepfaked. Instead, it was emphasized that offline, personal communication is the only trustworthy option.

However, in the discussions, many claimed that these misinformation and trust aspects that come with deepfakes are nothing new. A frequent example raised was Photoshop, which is an image editing application. They argue that distrust already exists and will move from just disregarding photos to videos and audio. Some users even saw this growing awareness and critical thinking around deepfake as a silver lining. However, as stated by other users, this still disallows two further communication forms (audio and video) that earlier possessed some trust. Simultaneously, users still saw Photoshopped pictures of people as an issue.

Another important concern for trust was highlighted in that the problem not only persists in that deepfake can be regarded as real, but that genuine content can be dismissed as fake. Which involved serious implications, such as dismissing truthful video evidence as unreliable. One example raised was whitewashing, where deepfakes also provide a tool for deniability, particularly for individuals in positions of power. Compromising evidence could be dismissed as fabricated, allowing individuals to evade responsibility for their actions.

Further concerns also surrounded the growing availability and future of the technology. Many users highlighted the ease of creating deepfakes without any prior knowledge. Likewise, in the few deepfake tutorial posts that were included in the analysis, the comment section also expressed how easy it was to follow. This accessibility and easy creation of deepfakes also prompted different concerns about future scenarios. For instance, fear was expressed that the technology might be able to give you anything you want to whoever you want on demand, just from a single picture and an easy description. While both implicating consent and the risk of misinformation, an emphasis was placed on the fact that trust in each other is also affected.

4.1.4. Abuse and Harmful Use

While abuse and harmful use can be seen as the bigger picture and a more abstract theme of the concerns above, this theme especially highlights the misuse of technology. An important distinction in the discourse is between good and bad use. Many users highlighted that it is not the technology at fault, but the people who misuse it who should be responsible. Several positive areas were discussed around deepfakes, such as entertainment like memes, satire, and art. The reaction of these circulating entertaining deepfaked videos further highlighted the enjoyment of these types of content. However, many of the comments in the overall discourse also included a fear of technology, especially in the wrong hands. While deepfakes nature is implied to make fake news and propaganda more accessible, it is dependent on the person who uses it.

A consequence of this distinction was discussed around “hidden” or “disclaimed” deepfakes. The former was referred to content that was not stated or recognizable as deepfaked, and the latter when content was explicitly stated or evident as deepfaked. While disclaimed deepfakes were seen as having potential positive use in cases such as entertainment, hidden deepfakes might be used for earlier-stated concerns such as for deception and framing.

The ethical concern emerged in this theme was that the bad use might outweigh the good, making a potential life improving technology shed a negative light. There were controversies about whether a ban, due to overall concerns on deepfakes, should take place and in many discussions this was spun up to create an ethical dilemma.

5. Discussion

This section discusses the study's results. Initially, the abstract findings of the study are outlined and compared with previous research, positioning the insights in the field of deepfake ethics. This comprehensive overview sets the stage for the subsequent discussion, where AI ethical principles are applied and examined in relation to the findings.

5.1. Discussion of Results

The result, through the identified themes of *Consent and Autonomy*, *Misinformation*, *Trust*, and *Abuse and Harmful Use*, explore the deeper societal and personal implications surrounding deepfakes. These themes collectively paint a picture of the erosion of personal boundaries, the increasingly complex challenge of discerning reality in the digital media, and the evolving nature of distrust. The concern of consent and autonomy, for instance, relates to the misuse and unauthorized use of one's digital persona. Misinformation, on the other hand, highlights the consequences of deepfakes deceptive nature. Trust reflects the concerns on the societal shift towards skepticism, while abuse and harmful use reveal the dual nature of technological advancements, as tools for both creation and exploitation.

Evidently, the exploration of social media discourse surrounding deepfakes unveils a range of ethical concerns. While these themes themselves resonate with existing research and findings of scholars like Mahmud and Sharmin (2021), Kietzmann et al. (2020), and Karnouskos (2020), they also introduce unique perspectives. The study brings new important perspectives within these more abstract concerns, emphasizing personal implications and insights into societal consequences.

Firstly, one distinctive aspect of the discourse was the focus on various actors affected by deepfakes. In line with Diakopoulou et al. (2019) and Kietzmann et al. (2020), who categorized groups harmed by deepfakes, the analysis also identified concerns spanning from individuals to broader societal actors. These conversations, however, goes beyond concerning those harmed by deepfakes to include those who might use them harmfully, and explores the personal implications of these actors' interaction with the ethical challenges posed by deepfakes. The insights show a varied landscape: viewers and listeners struggle with discerning truth from fiction, creators have to navigate the ethical implications of their creations, subjects deal with the repercussions of having their likeness used without consent, and societies face the growing risk of eroded trust in digital media.

While these actors in the deepfake context still to some degree are discussed in existing research, there are certain aspects of these groups that the concerns from social media discourse shed new light on. For instance, the conversations regarding viewers and listeners of deepfakes gave an important attention to vulnerable groups that are more prone to be harmed and deceived from directed misinformation. Further, the analysis also emphasized a broader concern on deepfake subjects. While research like that of Kietzmann et al. (2020) acknowledges that anyone can fall subject to non-consensual deepfakes, prior studies have often been focused on politicians and celebrities, particularly in the context of political campaigns and pornography. In contrast, the findings reveal a more widespread concern, indicating that ordinary social media users are increasingly apprehensive about being

deepfaked. This concern can be seen in a reluctance to post on social media platforms, highlighting a significant impact on personal autonomy and trust in digital media.

Moreover, this growing distrust also reveals a more profound societal concern than previously explored in existing research. The movement towards skepticism and the emergence of a "post-truth" society goes beyond the context from prior discussions of political trust or journalistic integrity by researchers like Karnouskos (2020) and Diakopoulou et al. (2019). The hesitation to post online and the preference for offline communication indicate that this distrust impacts not just the consumption of digital content such as journalism, but also what we post. The discourse further reveals how distrust can be exploited, for instance, in whitewashing, adding another layer of complexity to the issue. Going deeper, concerns arise about using deepfakes for personal alterations, such as appearing younger, challenging the trustworthiness of self-posted content. The future concern that anyone might be able to create convincing deepfakes with just a photo and a click further aggravate these trust issues, posing a significant challenge to societal trust as a whole.

When examining these concerns, it also becomes evident that they are not isolated problems but deeply interconnected. Misinformation issues feed into consent and autonomy, as seen in cases like the alleged Mr. Beast scam, consent is often disregarded in the pursuit of manipulated truth which consequently harms the autonomy of the subject. Similarly, the growing skepticism towards digital content, fueled by the rise of deepfakes and misinformation, undermines trust in digital media. This skepticism also affects how individuals autonomy to use their own likeness online.

Evidently, the exploration of social media discourse on deepfakes shows a complex interplay of personal and societal implications by this technology. There are multiple unique and profound insights and concerns within these themes to add to the comprehensive understanding of the ethical landscape surrounding deepfakes. While a big part of the findings resonate with existing research, the discourse highlights a deeper meaning and consequences of these implications and concerns.

5.2. Results in Relation to AI-Ethical Principles

When expanding the discussion beyond the specific concerns about deepfakes to examine how these identified aspects fit within the broader field of AI ethics, it becomes clear that deepfakes intersect with multiple established AI ethical principles, as outlined in Jobin et al.'s (2019) overview. This parallel not only highlight how the social and personal implications and values identified in the discourse are reflected and interpreted within the wider field of AI ethics, but also how these generalized principles are applied or diverged in the specific context of deepfakes

Seen to principles in Jobin et al.'s (2019) study, the most notable one in relation to the findings of this study is possibly non-maleficence: that AI technologies should not harm individuals or society. As identified in the result, the concerns encompassed potential harm to viewers/ listeners, subjects, and society at large. For instance, being deepfaked in pornographic content highlights the individual harm, and the larger effects on trust demonstrates the sharm on society. However, the principle of non-maleficence in the context

of deepfakes captures the ethical implications in a very generalized way. Looking deeper into Jobin et al.'s overview, multiple more of the principles can be found grounded in the themes that arose.

Firstly, regarding the theme of consent and autonomy, this concern can be seen to significantly relate to the privacy principle. Privacy, as discussed by Jobin et al. (2019), focuses on protecting personal data from unauthorized access and misuse. In the context of deepfakes, however, 'personal data' extends to include one's likeness. This encompasses not only the images, videos, and audios collected for the creation of deepfakes, but also the digital representation of an individual in its entirety. Despite users possibly not expecting complete privacy for publicly shared images, which may be used to create deepfakes, there's a clear expression of concern over their representation being unauthorized misuse. The divergence of this principle can be seen in pornographic deepfakes like the case of Atrioc, when the deepfaked streamers likeness was non-consensually used for profit.

Further, this misuse also touches on the principle of dignity, which Jobin et al. (2019) address as upholding human rights and individual values. The social media discourse revealed their remarks on the dehumanizing nature of deepfakes, that individuals are reduced to mere objects. This can be seen violating both the individual values found in dignity, but also autonomy which is a principle itself in the authors overview. Moreover, the concern of one's likeness being used for deepfakes unknowingly also calls upon the principle of transparency in AI usage. Jobin et al. also emphasize transparency's role in building trust, which in this study is evident in the users' reluctance to post content on social media due to the growing distrust to be non-consensually and unwitting deepfake creations.

In the case of misinformation, the principle of transparency is also an important aspect. However, from the perspective of this theme it does not regard how data or likeness are used. Rather, it regards how the deepfake content itself lacks transparency due to its realistic nature, especially when not disclaimed as manipulated. Similarly to how organizations that use AI systems for decision-making processes should disclaim the logic behind their conclusions, deepfakes are also concerned about clarity of their nature.

Additionally, in the context of misinformation, the principle of non-maleficence becomes particularly relevant again. Multiple instances were discussed that negatively impact both individuals and society through means such as fake news and framing. This also includes the concern for reputational harm in cases such as revenge porn, which also correlates to the aspects of dignity and autonomy. These concerned instances reveal several possibilities of deepfakes to violate these different notions.

Following the themes of consent, autonomy and misinformation and the challenges they pose to these principles, trust becomes an apparent issue within the context of deepfakes. This theme should be correlated to the principle in Jobin et al.'s (2019) framework of the same name. However, the concern around trust identified in this study does not consider the trust of AI in the technical aspect. To further clarify, in the authors overview, the trust principle can be exemplified as how algorithms are trusted to make correct decisions or if there is excessive trust in the decision. In the deepfake context, this could translate to be seen as trusting the technology to create high quality realistic deepfakes and imitations. However, the case of the concerns in the discourse highlight the distrust of the technology being transparent and used

for beneficial reasons, rather than the technology itself. As such, this theme instead can be seen also concerning the principles of transparency and non-maleficence. The perspective of what constitutes “beneficial” use in these discourses also correlates to if there is respect for the principles of one's privacy, autonomy and dignity.

Lastly, the abstract theme of abuse and harmful use encompasses a range of principles, including beneficence, responsibility, transparency, freedom, and autonomy. This theme, similarly to the non-maleficence position for societal and personal aspects in the overview of Jobin et al. (2019), can be seen to capture the widespread ethical implications and concerns surrounding deepfakes in the social media discourses. However, the perception around this theme also highlights how these principles are balanced in the context of deepfakes. For instance, some users remarked on the positive social outcomes of deepfakes, suggesting their alignment with beneficence. This can be seen in the case of artistic and entertainment purposes, where it can be a form of autonomy and expression. However, the violation and lack of transparency and potential harm associated with deepfakes highlight the ethical dilemmas in their use, questioning the balance between their beneficial and maleficent impacts.

5.3 Assessment and Reflection of Method and Results

The evaluation of both the research process and consequently its outcomes is a critical step. Seen to the methodology of passive approaches in netnography, Costello et al. (2017) highlights that studies incorporating this method often neglects such an assessment. Trying to set an example, this study has made an effort in reflecting deeply on engagement, responsibility, and co-creation of knowledge (aspects underlined by Costello et al.) in the presentation of the method. Yet, in the light of the principles of trustworthiness in qualitative studies, outlined by scholars such as Nowell et al. (2017), there are further criterias that require assessment.

The primary aim was to understand prevailing concerns about deepfakes in social media discourse, focusing on Reddit and YouTube. This purposive selection, though necessary, introduced biases and especially raises concerns regarding its transferability of the findings to the broader social media domain. Transferability in qualitative research, as described by Nowell et al. (2017), refers to the generalizability of inquiry, providing a thick description so others can judge the applicability of findings to their contexts. Following this, it should be noted that the selection may overlook certain social media spaces, groups, and demographics. Even within the chosen platforms, there is a risk that less popular or mainstream viewpoints might be excluded. Conversely, this study sought to identify the most prominent concerns, which inherently align with mainstream and popular viewpoints. The study's iterative and saturative approach suggests that the identified prominent concerns should be transferable to the discourse on this platform. However, it's important to acknowledge that not all social media users engage with or are even aware of deepfakes or the discussion surrounding it, which might set limitations in this study's representation of the users perception of these platforms' user base as a whole.

When looking past these platforms, it is also important to recognize that different platforms might encapture other demographics. Platforms might encompass a different user base with

cultural and social differences. If there is an evident difference in these aspects, it might contain other viewpoints, especially on the deepfakes circulating in their respective spaces. However, according to the demographics of Youtube and Reddit, they consist of significant user bases, with youtube even having around 80% of all online users active on their platform (Shepherd, 2024). When examining these platforms, it can be argued that they give a comprehensive and representative view of social media discourse.

Going even further than the stated purpose and looking at the population as a whole, social media discourse and perceptions might paint another picture. As Dixon (2023d) points out, social media users constitute approximately 61% of the global population. This leaves a significant portion of the population underrepresented or entirely absent in these digital discussions, making this transferability less applicable.

When discussing trustworthiness in qualitative research, and especially with thematic analysis, two important aspects are also credibility and confirmability. The former underlines that the respondents view should be correctly represented, the latter underlines how researchers' interpretations clearly demonstrate that it is derived from the data (Nowell et al., 2017). Due to the latent approach in the thematic analysis, and the ethical measure to not include citations or sharing the data, both the credibility and confirmability can be seen lacking transparency. Seeking consent from participants to share this data, although time consuming, could demonstrate these notions further. Moreover, Nowell et al. (2017) recommend member checking, letting participants respond to the interpretations accuracy and resonance, which could benefit these aspects in this study. Nevertheless, the study's detailed presentation on its process and activities can be seen following the criteria of dependability, meaning being logical, traceable and clearly documented (Nowell et al., 2017). This allows for repeatability, which can be seen migates the cost of the lack of confirmability.

Lastly, it should be highlighted surrounding the overall method used, the debated passive approach of netnography. As suggested by Kozinets (2010), a passive observer in research might not attain profound cultural comprehension, which could result in interpretations that are less insightful. Such an approach may lead to analyses that are superficial and descriptive, potentially missing the deeper cultural meanings embedded within the data. Seen again to the notion of credibility, and which Nowell et al. (2017) also recommend an active engagement. Even though the choice behind this approach had motives, a participatory observation might enhance the findings.

6. Conclusion

This study set out to explore the ethical landscape of deepfakes through the discussions and perceptions of deepfakes on social media platforms, identifying prominent concerns regarding personal and societal values and implications in this public discourse. Further, it seeked to bridge the theoretical aspects of the broad AI ethical landscape with the context of deepfakes, by investigating how these concerns can be interpreted through established AI ethical principles and simultaneously reveal how these principles might be violated in the light of these concerns.

Seen to the first research question, the findings from the discourse revealed several prominent concerns, ranging from the themes of consent, autonomy, truth, and trust, as well as abuse and harmful use. These challenges are not just theoretical constructs but expressed real-world impacts on individuals and society, giving insights on the personal consequences when being a subject of deepfakes, consuming deepfake content, to the societal effect it has on digital media and distrust.

The larger societal implications of deepfakes found centered around the ethical concern of this eroding trust was a particularly profound and unique insight in social media.The conversations on the platforms revealed a perspective on how digital media, communication, and societal interactions can see a shift in the light of deepfakes. This change manifests as varied forms of distrust: on what we dare to post, what we see and hear on digital media, and ultimately even on the trust in each other. The reluctance to share on digital platforms follows the concern for personal ramifications of non-consensual use of one’s likeness in deepfakes , and the infringement it can have on individual autonomy. The skepticism towards the authenticity of the content we consume online can be found significantly impacted by the misinformation propagated by deepfakes, which complicates the process of discerning reality. Lastly, the trust on eachother can be seen in the overall concern that people can use deepfakes for abuse and harmful use.

Moving to the second research question, when interpreting established AI ethical principles within these concerns, it also becomes evident that deepfakes possess more unique ethical implications than these generalized principles manifest. For instance, the notion of trust, which is an established principle itself, is a far more complex and entangled societal issue in the context of deepfakes. Further, the principle of privacy reveals how its focus on “personal data” might not fully capture the perspective on the consensual use of one's digital likeness and representation.

Despite how some concerns might be difficult to interpret, the societal and personal implications on deepfakes in the findings of this study still indicates instances when these principles might be violated in this emerging technology. The often non-consensual use of deepfakes questions how dignity can be upheld, the notion of misinformation and the deceiving nature of deepfakes question the principle of transparency and the distrust reveals potential harm to society and individuals, diverging the principle non-maleficence.

The concerns raised in this study offers a deeper understanding into the implications surrounding deepfakes. By highlighting the societal and personal values of a public response of deepfakes, it can contribute to the ongoing conversation about AI ethics. Further, it takes

action and gives knowledge to the transition from overgeneralized AI ethical standards to more focused, situational ethics that take into account the specificities affects of deepfake technology and its societal implications. By delving into the social media discourse, it reveals invaluable insights these values at play, which are critical for informing policy, regulation, and the responsible development of AI technologies.

6.1 Future Research

As deepfake and general AI technologies rapidly advance, it is essential that ethical discourse, that considers societal responses, evolve correspondingly. This ensures that they remain relevant and effective in tackling the challenges presented by these continuously developing technologies. Deepfake technology is already finding new applications in areas such as digital "AI Companions". Here, AI chatbots are being designed to use celebrity likenesses and personalities that users can interact with, integrating deepfake technology to send multimedia messages during these conversations.

Expanding research beyond social media discourse to encompass broader societal discussions could also be beneficial. As earlier stated, about 61% of the global population are social media users, leaving a significant demographic underrepresented in digital dialogues (Dixon, 2023d). While the concerns identified in the study primarily affect social media users, they are not exclusive to them. Deepfake implications still impact and concern those outside the social media environment. For instance, you don't have to be on social media to become the subject of a harmful deepfake. Further, deepfakes are seeing application in other areas, such as traditional media with deepfaked news anchors. As such, incorporating diverse ‘offline’ research methods, like interviews, could better represent these underrepresented groups.

Moreover, going beyond just the prominent concern and investigating a more comprehensive view of concerns could also inform less prevalent aspects of deepfakes. For instance, a specific application of deepfakes, which might not be as concerning or not encompass as much awareness, might still be a significant risk that needs to be assessed and measured.

Lastly, while this study contributes descriptively to the discourse on deepfake and AI ethics, there's a need for research on converting these concerns and insights into prescriptive actions. Implementing ethical principles, frameworks, and guidelines, as discussed by Hagendorff (2019) and Munn (2023), remains challenging. For social media platforms, specifically, research should focus on how to effectively mitigate the harm indicated in these concerns. Actions against misinformation, including deceptive deepfakes, are already being taken, as seen with platform X’s recent implementation of “community notes” (Elliott & Gilbert, 2023). Such initiatives are crucial steps towards fostering trust and reducing deception in digital content. Future research should aim to build upon these foundations, exploring effective strategies and solutions for the ethical challenges posed by deepfakes.

7. References

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The state of deepfakes: Landscape, threats, and impact. Amsterdam: *Deeptrace, 27*. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

Benamara, F., Inkpen, D., & Taboada, M. (2018). Introduction to the Special Issue on Language in Social Media: Exploiting Discourse and Other Contextual Information. Computational Linguistics. *Computational Linguistics, 44(4)*, pp. 663–681.https://doi.org/10.1162/coli_a_00333

Berg, M. (2015). *Netnografi - att forska om och med internet*. Lund: Studentlitteratur.

Bertilsson, J. (2014). Netnografi - en metod för att studera internetbaserad kommunikation. In J. Eksell, & Å. Thelander (Eds.), *Kvalitativa metoder i strategisk kommunikation* (pp. 111-126). Studentlitteratur.

Bryman, A. (2018). *Samhällsvetenskapliga metoder. (3. Edition)*. Stockholm: Liber AB.

Boynton, E. (Producent). (13 december 2023). *Behind-the-scenes at an ethical porn shoot*. [Video]. BBC. https://www.bbc.com/reel/video/p0gytpbk/behind-the-scenes-at-an-ethical-porn-shoot

BuzzFeedVideo. (17 april 2018). *You Won't Believe What Obama Says In This Video!* [Video]. YouTube. https://www.youtube.com/watch?v=cQ54GDm1eL0&ab_channel=BuzzFeedVideo

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, vol 3(2)*, pp. 77–101. https://doi.org/10.1191/1478088706qp063oa

Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. (2021). *Deepfake: An Overview. Lecture Notes in Networks and Systems, 203.* https://doi.org/10.1007/978-981-16-0733-2_39

Chan-Olmsted, S. M. (2019). A Review of Artificial Intelligence Adoptions in the Media Industry. *International Journal on Media Management, 21(3–4)*, pp. 193-215. https://doi.org/10.1080/14241277.2019.1695619

Chen, P. J. (2013). Social media. In *Australian Politics in a Digital Age* (pp. 69–112). ANU Press. http://www.jstor.org/stable/j.ctt2jbkkn.11

Cole, S. (26 august 2019). A Site Faking Jordan Peterson's Voice Shuts Down After Peterson Decries Deepfakes.*Vice*. https://www.vice.com/en/article/43kwgb/not-jordan-peterson-voice-generator-shut-down-deepfakes

Cole, S. (31 january 2023). Deepfake Porn Creator Deletes Internet Presence After Tearful 'Atrioc' Apology. *Vice*. https://www.vice.com/en/article/jgp7ky/atrioc-deepfake-porn-apology

Costello, L., McDermott, M. L., & Wallace, R. (2017). Netnography: Range of Practices, Misperceptions, and Missed Opportunities. *International Journal of Qualitative Methods, 16(1)*. https://doi.org/10.1177/1609406917700647

Deeptomcruise. (26 december 2022). *When I dance, I dance with my hands* [Video]. TikTok. https://www.tiktok.com/@deeptomcruise/video/7181490100314885382

Dixit, P. (1 november 2023). Rashmika Mandanna deepfake: Delhi Police initiate inquiry, seek URL of video from Meta. *Business Today*. https://www.businesstoday.in/technology/news/story/rashmika-mandanna-deepfake-delhi-police-initiates-inquiry-seeks-url-of-video-from-meta-405456-2023-11-11#:~:text=The%20Delhi%20Police%20has%20formally,in%20connection%20with%20the%20 incident.

Dixon, S. J. (13 september 2022a). *Reddit - Statistics & Facts*. Statista. https://www.statista.com/topics/5672/reddit/#topicOverview

Dixon, S. J. (15 december 2023b). *Distribution of Reddit users worldwide as of 3rd quarter 2022*, by gender. Statista. https://www.statista.com/statistics/1255182/distribution-of-users-on-reddit-worldwide-gender/

Dixon, S. J. (2 august 2022c). *Percentage of U.S. adults who use Reddit as of February 2021*, by age group. Statista. https://www.statista.com/statistics/261766/share-of-us-internet-users-who-use-reddit-by-age-group/

Dixon, S. J. (10 januari 2024d). *Social media - Statistics & Facts Worldwide*. Statista. https://www.statista.com/topics/1164/social-networks/#topicOverview

Etik. (April 18, 2018). Vetenskapsrådet. https://www.vr.se/uppdrag/etik.html

Elliott, V., & Gilbert, D. (17 october 2023). Elon Musk's Main Tool for Fighting Disinformation on X Is Making the Problem Worse, Insiders Claim. *Wired*. https://www.wired.com/story/x-community-notes-disinformation/

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schäfer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. s. *Minds and machines (Dordrecht), 28 (4),* pp. 689-707. https://doi.org/10.1007/s11023-018-9482-5

Franzke, A. S., Bechmann, A., Zimmer, M., Ess, C., & Association of Internet Researchers. (2020). *Internet Research: Ethical Guidelines 3.0*. https://aoir.org/reports/ethics3.pdf

Guarnera, L., Giudice, O., & Battiato, S. (2020). Fighting Deepfake by Exposing the Convolutional Traces on Images. *IEEE Access, 8,* pp. 165085-165098. https://doi.org/10.1109/access.2020.3023037

Taherdoost, H. (2016). Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. *International Journal of Academic Research in Management, 5,* pp. 18-27. https://doi.org/10.2139/ssrn.3205035

Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine, 292*. https://doi.org/10.1016/j.socscimed.2021.114523

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and machines (Dordrecht), 30(1),* pp. 99-120. https://doi.org/10.1007/s11023-020-09517-8

Diakopoulos, N., & Johnson, D. G. (2019). Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections. *New media & society, 23(7)*, pp. 2072–2098. https://doi.org/10.2139/ssrn.3474183

Jobin, A., Ienca, M & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, pp. 389-399. https://doi.org/10.1038/s42256-019-0088-2

Karnouskos, S. (2020,). Artificial Intelligence in Digital Media: The Era of Deepfakes. *IEEE Transactions on Technology and Society, 1(3)*, pp. 138-147. https://doi.org/10.1109/tts.2020.3001312

Khanjani, Z., Watson, G., & Janeja, V. P. (2023). Audio deepfakes: A survey. *Frontiers in Big Data*. https://doi.org/10.3389/fdata.2022.1001063

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

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., & Theobalt, C. (2018). Deep video portraits. *ACM Transactions on Graphics, 37(4)*, pp. 1-14. https://doi.org/10.1145/3197517.3201283

Kozinets, R. V. (2010) *Netnography : doing ethnographic research online*. Los Angeles, Calif: SAGE.

Kozinets, R. V. (2015) *Netnography : redefined*. 2nd edition. Thousand Oaks, CA: Sage Publications Ltd.

Kudhail, B. P. (12 october 2023). Could an AI-created profile picture help you get a job? *BBC News*. https://www.bbc.com/news/business-67054382

Kou, Y., Kow, Y. M., Gui, X., & Cheng, W. (2017). One Social Movement, Two Social Media Sites: A Comparative Study of Public Discourses. *Computer Supported Cooperative Work, 26(4–6)*, pp. 807-836. https://doi.org/10.1007/s10606-017-9284-y

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P., & Schroeder, D. T. (2021). Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes. *Frontiers in Communication*. https://doi.org/10.3389/fcomm.2021.632317

Leibowicz, C. R., McGregor, S., & Ovadya, A. (2021). The Deepfake Detection Dilemma. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3461702.3462584

Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z., Prasad, C., & Palaniappan, K. (2020). Deepfake Video Detection Based on Spatial, Spectral, and Temporal Inconsistencies Using Multimodal Deep Learning. *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. https://doi.org/10.1109/aipr50011.2020.9425167

Liao, S. M. (2020r). A Short Introduction to the Ethics of Artificial Intelligence. *Ethics of Artificial Intelligence*, pp. 1-42. https://doi.org/10.1093/oso/9780190905033.003.0001

Maddocks, S. (2020) 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. *Porn studies (Abingdon, UK)*. [Online] ahead-of-print (ahead-of-print), pp. 1-9.

Mahmud, B. U., & Sharmin, A. (2020). Deep Insights of Deepfake Technology : A Review. https://www.researchgate.net/publication/351300442_Deep_Insights_of_Deepfake_Technology_A_Review

Morsello, B. (2017). The datafied society, studying culture through data. Information. *Communication & Society, 20(12)*, pp. 1824-1826. https://doi.org/10.1080/1369118x.2017.1366541

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2022.113368

Munn, L. (2022). The uselessness of AI ethics. *Ai and ethics, 3(3),* pp. 869-877. https://doi.org/10.1007/s43681-022-00209-w

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

Neubaum, G., & Krämer, N. C. (2016). Monitoring the Opinion of the Crowd: Psychological Mechanisms Underlying Public Opinion Perceptions on Social Media. *Media Psychology, 20(3)*, pp. 502-531. https://doi.org/10.1080/15213269.2016.1211539

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q. V., & Nguyen, C. M. (2022). Deep Learning for Deepfakes Creation and Detection: A Survey. *Computer vision and image understanding*. https://doi.org/10.2139/ssrn.4030341

Nirkin, Y., Keller, Y., & Hassner, T. (2019). FSGAN: Subject Agnostic Face Swapping and Reenactment. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 7183-7192. https://doi.org/10.1109/iccv.2019.00728

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods, 16(1)*, pp. 1-13. https://doi.org/10.1177/1609406917733847

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021, April). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society, 7(2)*. https://doi.org/10.1177/20563051211019004

Regeringskansliet, R. och. (2015, April 27). Vetenskapsrådet. *Regeringskansliet*. https://www.regeringen.se/myndigheter-med-flera/vetenskapsradet/

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2017). Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity, 52(4),* pp. 1893-1907. https://doi.org/10.1007/s11135-017-0574-8

Sebastian, B. M. (2023, November 7). Rashmika Mandanna calls for action against "scary" deepfake video. *BBC News*. https://www.bbc.com/news/world-asia-india-67305557

Shepherd. J. (5 january 2024). 23 Essential YouTube Statistics You Need to Know in 2024. *The Social Shepard*. https://thesocialshepherd.com/blog/youtube-statistics

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics. *Journal of Database Management, 31(2)*, pp. 74-87. https://doi.org/10.4018/jdm.2020040105

Swedish Research Council, V. (2002). Forskningsetiska principer inom humanistisk-samhällsvetenskaplig forskning. https://www.vr.se/analys/rapporter/vara-rapporter/2002-01-08-forskningsetiska-principer-inom-humanistisk-samhallsvetenskaplig-forskning.html

Verdier, H. (1 October 2018). Can Porn Be Ethical? review – a refreshing debate with no groaning sound effects. *The Guardian*. https://www.theguardian.com/tv-and-radio/2015/oct/01/can-porn-be-ethical-debate-sex-industry-working-conditions

van der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: regulatory challenges for the synthetic society. *Computer Law & Security Review, 46*. https://doi.org/10.1016/j.clsr.2022.105716

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *TIM Review, 9(11)*, pp. 39-52. https://timreview.ca/article/1282

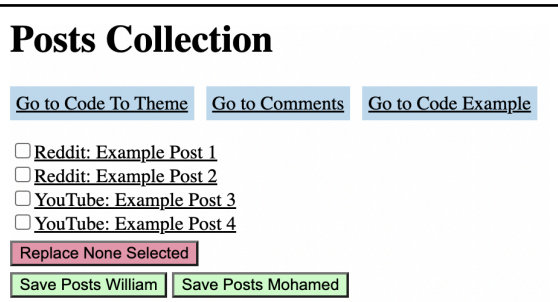
Widder, D. G., Nafus, D., Dabbish, L., & Herbsleb, J. (2022). Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. *2022 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3531146.3533779

Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A Survey on Deepfake Video Detection. *IET Biometrics, 10(6)*, pp. 607-624. https://doi.org/10.1049/bme2.12031

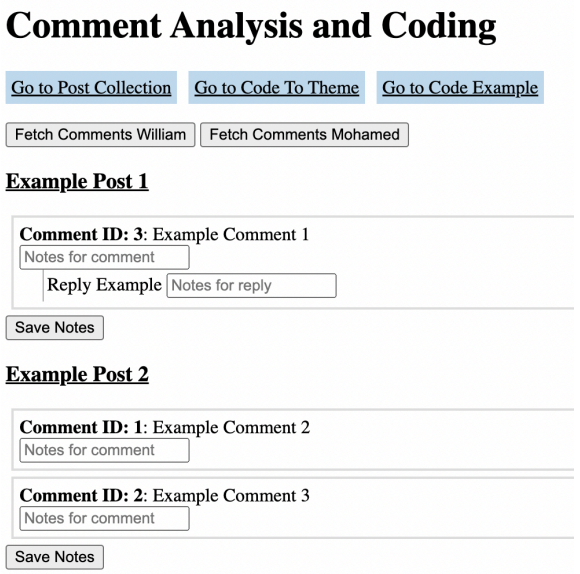
8. Appendix A: Data Collection and Analysis Artefact

Post retrieval

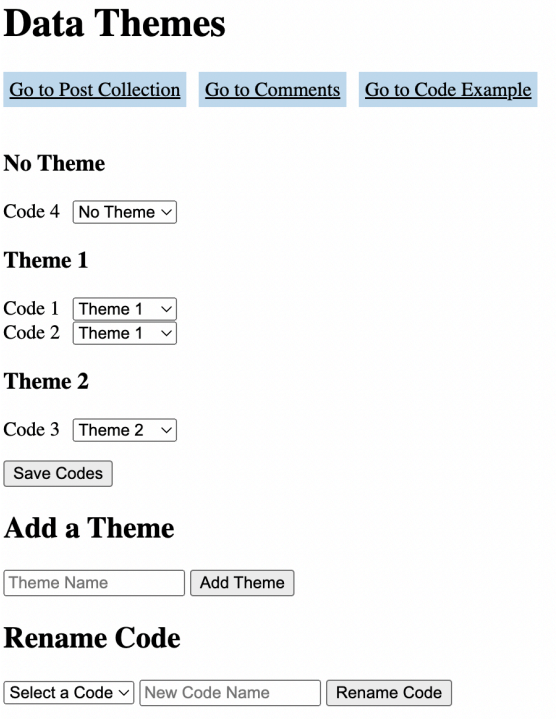
Selection and replacement

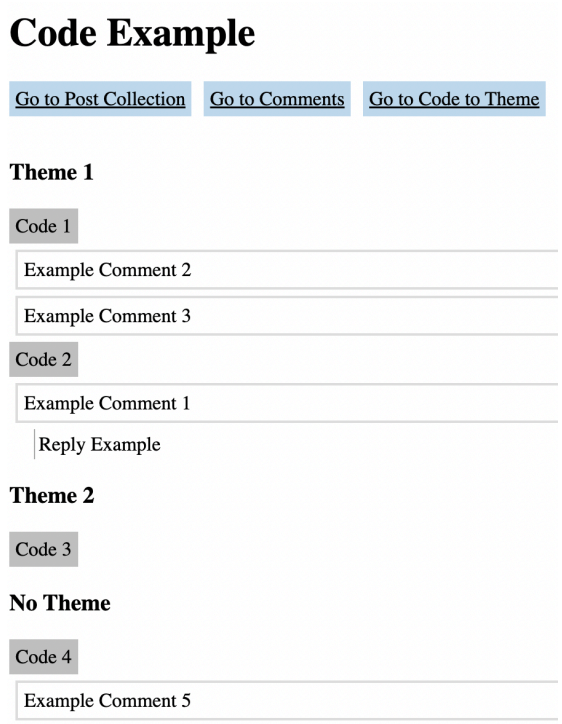


Display, note and code the comment section



Create and assign themes, rename codes




Code Example
Go to Post Collection
Go to Comments
Go to Code to Theme
Theme 1
Code 1
Example Comment 2
Example Comment 3
Code 2
Example Comment 1
Reply Example
Theme 2
Code 3
No Theme
Code 4
Example Comment 5