ABSTRACT

Title of Thesis: CREEPY OR COOL? AN EXPLORATION OF NON-MALICIOUS DEEPFAKES THROUGH ANALYSIS OF TWO CASE STUDIES

Keaunna Cleveland
Master of Science in
Human-Computer Interaction, 2022

Thesis Directed By: Dr. Katie Shilton, Associate Professor
College of Information Studies

Several studies have examined the harms associated with the development of deepfake technology and its use by malicious actors, but less research has been devoted to deepfakes created by non-malicious creators and the ways people react to deepfakes developed without malicious intent. This study attempts to close this research gap through the exploration of two case studies that demonstrate non-malicious deepfake use on Instagram and Twitter. Using sensemaking, privacy as contextual integrity, and audience theory to guide the analysis of publicly available posts, tweets, and records, this study examines how people interact with and react to non-malicious deepfakes online. Building on these findings, this thesis suggests how social media platforms might integrate signifiers in their design that afford sensemaking for those interacting with deepfake technology and discusses how ethical frameworks and practices

from values-oriented design and value-based engineering in design may help guide creators as they develop deepfake technology videos and applications for non-malicious purposes.

CREEPY OR COOL?
AN EXPLORATION OF NON-MALICIOUS DEEPFAKES THROUGH ANALYSIS OF TWO CASE STUDIES

by

Keaunna Cleveland

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2022

Advisory Committee:
Associate Professor Katie Shilton, Chair
Assistant Professor Daniel Greene
Associate Professor Jessica Vitak

Acknowledgments

I would like to express my sincerest gratitude to my advisor, Dr. Katie Shilton, who guided me throughout this project, and whose insight was invaluable to my work. I would also like to thank Dr. Daniel Greene and Dr. Jessica Vitak for their feedback, support, and guidance. Without my committee, this research would not have been possible.

I wish to acknowledge my fellow EViD lab members, who provided me with feedback, suggestions, and critique. I would also like to show my deep appreciation to my writing group, whose members gave me the space to talk through new ideas, as well as to Kausalya Ganesh, who acted as my informal mentor.

I would like to thank my sister, Miyaunna Vásquez Sharree, who was always there to listen as I went through the difficult process of getting my ideas on paper.

Finally, thank you to the staff, faculty, and students in the HCIM program and iSchool who provided me with both formal and informal support, feedback, and guidance.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In 2017, Reddit user *Deepfakes* posted to the subreddit r/deepfakes what appeared to be videos of famous actresses and (female) celebrities performing in pornographic videos. Closer inspection revealed these videos to be inauthentic; *Deepfakes* created these videos by leveraging artificial intelligence and video effects techniques, superimposing images of famous women over the original faces in each video (Pérez Dasilva et al., 2021). The result was a collection of high-quality deepfake videos that looked convincingly real. These "deepfakes" (a combination of "deep learning" and "fake") represent the emergence of a new type of synthetic media in which one person's likeness is replaced with that of another person's; the purpose is to present one person as having said and done the actions of the other.

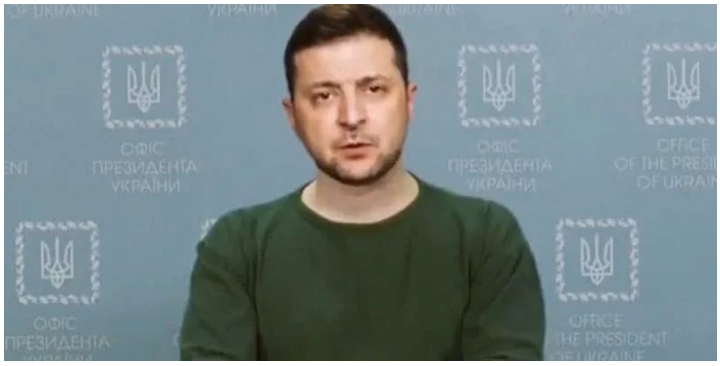
Awareness of deepfakes has since erupted in the academic and public consciousness, particularly as deepfakes have spread from Reddit to other online spaces and have crossed from nonconsensual pornography to political misinformation (Greengard, 2019). As deepfake popularity and use have grown, several studies have also sought to examine the harms and ethical dilemmas associated with the development of this technology and its use by malicious actors to create deceptive images, video, and audio (Meskys et al., 2020). Deepfake technology is also used in non-malicious ways, where the intent behind the creation of each video is less insidious (Meskys et al., 2020).

1.1 Motivation

Deepfake technology presents a unique dilemma. On one hand, this technology has been used (as in the case of nonconsensual and revenge pornography) to strip people of their agency, often with devastating and life-altering negative consequences (Maddocks, 2020). Further, the use of deepfake technology to spread *convincing* misinformation positions the technology as a dangerous weapon with far-reaching social and political ramifications (Greengard, 2019). A more recent example involved a deepfake created of Ukrainian President Volodymyr Zelenskyy: in the minute-long video, the fake attempts to convince Ukrainian soldiers to surrender to Russian forces**.**

Figure 1

Example of a Malicious Deepfake



Note. Screenshot from a video featuring a deepfake of Ukrainian President Volodymyr Zelenskyy (Allyn, 2022).

On the other hand, like with video and image editing software and technology, deepfake technology allows people to explore creative and technical possibilities that would otherwise be closed to them. Technologists, content creators, and others have certainly seized on the opportunity that deepfake technology presents for enhancing and expanding the scope and types of projects they can create. Some of the newer iterations of deepfake technology use have included recreations of famous artists like Salvador Dali, humorous mashups featuring Nicolas Cage and other actors, and a public awareness commercial about malaria featuring a deepfaked David Beckham "speaking nine languages" (Kerner & Risse, 2021; Mihailova, 2021). A popular YouTube video developed by Jordan Peele featured a deepfake President Obama warning of nefarious deepfake use.

Figure 2

Example of a Non-Malicious Deepfake

Note. Screenshot from a deepfake video of Barack Obama. (BuzzFeedVideo, 2018)

Given the relatively recent appearance of these non-malicious deepfakes, however, there is a lack

of research that has examined the overall landscape of its non-malicious use and even less research that attempts to explain the interactions and reactions people have as they work to make sense of deepfake technology and videos. Specifically, this study focuses on deepfake technology within two contexts: 1) the creation of deepfake videos for a specific audience (by a single creator, focusing on a specific target, and developed for entertainment purposes), and 2) the use of an app to create deepfake videos featuring family, friends, and historical features, developed by a company for its customers.

1.2 Research Questions

To address this research gap outlined in **Section 1.1**, this thesis will explore how deepfakes are developed and used for non-malicious (creative) purposes and explore reactions to these deepfakes on social media. For this study, "non-malicious purposes" refers to deepfakes where the *main intent* behind the creation of the deepfake is not to deceive, and the deception is explicitly revealed by the creator, either through the use of the word "deep" or by a clear indication that deepfake/synthetic technology is being used to create the video. **Specifically, the following research questions were addressed:**

RQ1. In what ways do deepfake creators characterize their use and development of deepfakes and deepfake technology on social media?

- **a.** How are these deepfakes developed and used?
- **b.** What interactions (if any) do creators have with those who consume these deepfakes or deepfake technology?

RQ2. In what ways do people react to non-malicious deepfakes on social media?

a. How can these reactions be characterized?
b. What contextual factors (if any) do people identify, mention, or imply as influencing their reactions?

1.3 Approach

To address the above research questions, I decided to focus on two popular uses of non-malicious deepfake technology use trending on Instagram, Twitter, and in online news reports during 2021 (January – October 2021). In the first case, I collected data on Instagram surrounding the *deeptomcruise* tag and deepfake videos featuring a deepfaked Tom Cruise created by a filmmaker and VFX creator anonymized as *df_creator.* For the second case, I collected data from Twitter, focusing on tweets surrounding the *#DeepNostalgia* and *#MyHeritage* hashtags and mentions, and examined replies and conversations driven by these tweets. The use of synthetic media (deepfakes) on video sharing (YouTube) and social media websites has been studied in the past but focusing on trending uses of deepfakes on Instagram and Twitter presents an opportunity to expand research on the use of deepfake technology in these spaces. This study also attempts to understand and explore the conversations and interactions people have amongst each other with respect to these deepfake videos.

1.4 Outline of Thesis

This thesis is structured as follows: **Chapter 2** is a review of the literature, related work, and background on deepfake technology. This also explored how Audience Theory, privacy as contextual inquiry, and sensemaking play a role in how people react to non-malicious deepfake

technology. **Chapter 3** presents methods and findings of Case Study 1, which examines *deeptomcruise* on Instagram**.** Analysis of reactions to these videos showed that people used information-seeking behavior to make sense of deepfake videos. **Chapter 4** presents methods and findings of Case Study 2, which examines the use of the #DeepNostalgia hashtag on Twitter. Findings showed that given the oft-personal nature of these deepfake videos (generated from images of family members and historical figures), responses and reactions often prompted information-giving behavior. **Chapter 5** discusses the findings from this research with respect to broader theories on sensemaking, privacy as contextual integrity, and audience. The findings suggest that people use a blend of sensemaking strategies to understand deepfakes. Further, many people are curious about the implications of ethical deepfake use, regardless of creator intent, and worry about future harms. I also discuss frameworks for ethical deepfake development and use. Finally, **Chapter 6** provides a summary of the cases, findings, and discussion.

Chapter 2: Literature Review

This review provides an overview of prior research on deepfake technology use, including the democratization of deepfake technology through the development of open-source tools and publicly available web and mobile applications. The review will also discuss prior research that considers the ethical and moral implications of deepfake technology development and use, and present emerging empirical research centered on public reactions to deepfakes in museums, public spaces, and on video-sharing and social media sites.

2.1 The Emergence of Deepfakes

While the concept of face-swapping has roots in other types of photo and video manipulation methods (Albahar & Almalki, 2019), rapid developments in artificial intelligence technology and machine learning algorithms have greatly improved the techniques used to create deepfakes (Karnouskos, 2020). Many early studies on deepfakes have centered on determining what exactly deepfakes are, developing classification and detection systems, and providing context behind deepfake creation (Vizoso et al., 2021).

2.1.1 Cases and Types

There have been many different approaches advanced to identify and detect deepfakes. According to Meskys et. al (2020), most deepfakes can be categorized within *cases*, namely: political, pornographic, commercial, and creative (Meskys et al., 2020). Political deepfakes were often created from speeches of politicians, news reports, and socially significant events (Meskys et al., 2020). Creative/original deep fakes center on parodying and *memeing* individuals (mostly famous actors), and commercial deepfakes are those produced by companies

for various advertising or promotional goals (Meskys et al., 2020). Pornographic deepfakes represent the most common type of deepfake, in which the faces of famous actresses and everyday women are superimposed over those of pornographic film actresses (Meskys et al., 2020). Deepfakes have also been categorized by *type.* Kietzmann et al. (2020) arranged deepfakes into four categories: photo (face/body swapping), audio (voice swapping, text-to-speech), video (face-swapping/morphing, and full-body puppetry), and audio-video, which uses a combination of the previously mentioned techniques**.**

2.1.2 Framing and Indexing

Some research has investigated reactions to deepfakes in terms of framing ("positive", "negative", "believable", "not believable") and indexing ("is authentic", "is a deepfake") of deepfakes has played a role in the way individuals react to them online, specifically in networked spaces like YouTube, Twitter, and Reddit. (Bode, 2021) analysis of the responses to deep fake videos on YouTube featuring Keanu Reeves showed that the context of the deepfake (that is, who created the deepfake, and their motivations), as well as the willingness of actors to create and/or correct the index impacted to the way people reacted to these videos (Bode, 2021). Removing the deepfake from its original context prevented others from correcting the index appropriately (Lee et al., 2021).

Lee et al. (2021) also examined the framing and indexing of deepfake videos by users online. Analyzing the comments sections of the top 10 most popular deepfakes found on YouTube (n=2689), their study showed that while half of the deepfake YouTube videos were framed positively, audience reactions were mostly neutral or negative (Lee et al., 2021). Further

investigation showed that the addition of commentary to the deepfake video made audiences more likely to find the video realistic (Lee et al., 2021).

2.1.3 Democratization of Deepfake Technology

Research has shown that deepfake technology has become more accessible to specialists and the general public via advances in video effects software, as well as the development of easily downloadable mobile and web applications (Kietzmann et al., 2020; Pu et al., 2021). The addition of powerful generative adversarial networks (GANs) and Auto Encoders(AEs) to open-source tools has also made developing deepfakes much simpler process for their creators (de Seta, 2021; Meskys et al., 2020; Pu et al., 2021; B. Zhang et al., 2020)

Analysis of a large dataset of online deep fake videos available (n=1,869) indicated that most of these deepfake creators come from two main groups: the first group consisted of content creators with specialized knowledge and tools available that allowed them to create large collections of deepfakes for specific audiences (Pu et al., 2021). The second group consisted of those who uploaded (and by implication, created) only one or two deepfake videos (Pu et al., 2021). Deepfakes developed by both groups were created using a combination of open-source tools, public applications, and other undetermined methods (Pu et al., 2021).

2.1.4 Popularity and Spread

Prior research suggests that non-consensual pornographic deepfakes are the most common form of deepfake online (Maddocks, 2020). These deepfakes are created not only from images of celebrities but also from images of everyday women who may be known or unknown to the

deepfake creator (Maddocks, 2020). Although many deepfakes are (non-consensual) pornographic deepfakes, attention and conversation around political deepfakes also remain popular (Pérez Dasilva et al., 2021). Political deepfakes are of great interest to people online and are among the types of deepfakes shared the most (Ahmed, 2021). Compared with people in Singapore, people in the United States were more likely to have had some exposure to deepfakes, had larger social networks, and had greater political interest (Ahmed, 2021). Their higher political interest coupled with an exposure to deepfakes led to more inadvertent sharing of deepfakes.

Research has shown that the most *talked-about* (and by implication, highly spread) deepfakes are political deepfakes, as well as those satirizing and memeing famous actors/actresses, politicians, athletes, and other well-known public figures (Pérez Dasilva et al., 2021). Journalists and the news media in general also hold a high degree of power over the conversation surrounding these political deepfakes, perhaps in particular because of the vocal nature of their objection to deepfakes and the potential harm they might cause (Pérez Dasilva et al., 2021). Journalists and news media organizations have taken a supervisory approach to regulating the spread of deepfakes, working with major social media platforms to categorize and label deepfakes (Vizoso et al., 2021).

2.2 Ethical, Philosophical, and Legal Landscape

As is the case with most new technology, malicious actors have exploited deepfake technology for nefarious purposes (Greengard, 2019). Previous research on deepfake technology has focused

on the threats of political and pornographic deepfakes and the harm these may cause to individuals, societies, and institutions.

2.2.1 The Ethical Dilemma

Much theoretical and philosophical research has focused on the ethical questions deepfakes raise. de Ruiter (2021) focused on three factors to determine if deepfakes are morally wrong, namely: 1) objections from the person or people who have been deepfaked, 2) whether the deepfake is successful in its deception, and 3) whether the deepfake was created with malicious intent (i.e., with the purpose to deceive). The study showed that although these deepfakes are not necessarily intrinsically morally wrong, they do present an ethical dilemma and are intrinsically morally suspect (de Ruiter, 2021). A study on the ethical considerations of non-consensual pornography also focused on the moral dilemma inherent in deepfake technology (Öhman, 2020).

2.2.2 The Potential Threat to Knowledge

The emergence of deepfakes has driven theoretical and philosophical research examining whether their existence contributes to an overall erosion of knowledge. According to Rini (2020) audio and video recordings are part of public discourse, called "testimony"', and truth-telling is promoted among those who give testimony because the nature of audio and video recordings provide a definitive record of truth that can be continuously referenced (Rini, 2020). Deepfakes, however, erode the ability of recordings to be acute correctors -- recordings can no longer correct the record because the record can always be disputed (Rini, 2020). This in turn degrades the public's willingness to give accurate testimony as there is no incentive to maintain the truth. In that respect, deepfakes were found to not simply be dangerous because they are grounded in

deception, but also because they regulate behavior away from truthful testimony, eroding knowledge overall (Rini, 2020).

Like Rini, Kerner and Risse (2021) also maintain that deepfake technology poses an epistemic threat. Deepfakes challenge epistemic *actorhood* (seeking or revealing information collectively or individually as objects or subjects) by reinforcing the spread of misleading information, reducing the ability of the actor to give accurate information, and promoting false information about actors. While acknowledging digital technology has always had epistemic challenges, their paper highlights how deepfakes continue to degrade the ability of video to maintain truthful testimony; deepfakes encourage a no-truth network of epistemic actors that cannot determine truth from falsehood and will eventually have no incentive or tools to do so (Kerner & Risse, 2021).

2.2.3 Harm and Misinformation

The online spread of misinformation – false or misleading information intentionally or unintentionally used to deceive people – has emerged as a major issue in recent years (Wu et al., 2019). Social media platforms like Twitter and Facebook remain key online spaces where many people are exposed to and believe misleading or false information (Allcott et al., 2019). Trusted sources of information (such as content creators) also often unintentionally spread misinformation, and under their endorsement (via likes and retweets), people ultimately accept this false information (Wu et al., 2019). Disinformation, a more malicious form of misinformation in which false content is generated or spread intentionally to mislead people, has not only led to increased political polarization but also a weakening of truthful information

sharing (Wu et al., 2019). What's more, misinformation has worked to erode the credibility of truthful actors (Wu et al., 2019).

Like misinformation and disinformation, the use of malicious deepfake technology (itself a form of misinformation) has caused both individual and social harm and presents a clear threat to political systems in much the same way as older forms of misinformation (Kerner & Risse, 2021). The review also showed that the potential for deepfakes to influence political events and social movements is high; there have been several examples since the emergence of deepfake technology of people (including political figures) fooled by a deepfake video (Westerlund, 2019).

2.2.5 Potential Benefits

Still, deepfake technology use has more recently moved beyond malicious creations, often to enhance or explore experiences in novel ways. Mihailova (2021) examined the nascent symbiotic relationship between deepfakes and art spaces with a focus on how deepfake technology can be rehabilitated through use in museums. Focusing on three case studies (the reanimation of Salvador Dalí at the Dalí Museum, the film Warriors, and a fake advertisement, Wearing Gillian), deepfakes were shown to be used not as the focal point of the art piece, but as a means to explore a deeper issue; in many cases, deepfake use underscored an inherent unease with synthetic technology itself and highlighted tensions between deception and consent (Mihailova, 2021). The study also showed that deepfakes have the potential to increase the ways artists and others can display creative expression, and positioned museums, galleries, and art spaces as the domain in which deepfakes can be safely explored. Other examinations of beneficial deepfake

use have centered on identity protection, obscuring the identity and background of those from vulnerable groups while providing a human-like face to humanize the experience for both the speaker and viewer (de Ruiter, 2021).

2.3 Theoretical Framework

Like their malicious counterparts, non-malicious deepfake videos exist at the intersection of artificial intelligence, computer vision, and digital and visual media, and as mentioned in Section 2.2.5 have moved beyond malicious misinformation, fake news, and non-consensual pornography, to memes, art, humor, commercial uses (like in advertisements), and creative apps. Their ability to be transmitted on social and digital media platforms has also allowed them to spread across a variety of audiences. This flexible nature and the way social media affords a specific type of interaction and communication has allowed non-malicious deepfakes to take on three distinct forms: as a message, interaction, or communication between creators and their audiences, as a medium to transmit information that holds context and conveys meaning, and as a type of digital innovation that people react to and make sense of.

Given the relatively recent appearance of non-malicious deepfakes on social media, there remains a lack of research investigating how people interact with and make sense of this type of media in its creative form. Prior research on sensemaking, privacy as contextual integrity, and audience theory, however, might provide a lens for approaching and understanding non-malicious deepfake technology use and reception in its forms.

2.3.1 Sensemaking

Sensemaking is both a task-based and information-seeking strategy (P. Zhang & Soergel, 2014). According to Klein and Moon (2006), sensemaking involves "a motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively" (Klein & Moon, p.71, 2006). This definition is in part based on Weick's theory of sensemaking, in which people collectively confront gaps in their individual and collective understanding of a situation, task, or scenario by gathering data, fitting this data into their current frame(s) of understanding, and rejecting old and new information that does not match the new, developing understanding (Weick, 1995). This process is often retrospective and used to make sense of and explain past behavior and actions (Weick, 1995). Subsequent research in human-computer interaction (particularly in computer-supported cooperative work) has also used sensemaking theory to examine and understand both the individual and collective decision-making behavior of people in a variety of contexts and scenarios, such as groups completing computing tasks, or people interacting and collaborating on social media (Pirolli & Russell, 2011). Zhang and Soergel's (2014) comprehensive model of sensemaking builds upon these models and research from organizational theory, human-computer interaction, cognitive science, learning theory, and library and information science to present a cognitive understanding of sensemaking:

> [S]ensemakers identify a problem, realize they need more information, and, through exploring or browsing or broad search, learn about what information they need to update their knowledge. (p.35)

By design, social media platforms allow people to transmit information (news, opinions, other media, etc.) to others, which can then be absorbed and discussed *between* those on the platform, through comments, and replies. According to Marwick and Boyd (2014): "Individuals contribute text, photos, and other content, and 'like,' 'favorite,' and comment on other people's content to both recognize and engage with others" (p. 1054). The designs of both Twitter and Instagram also afford these types of interactions between users (even replies to replies), and in this way, micro-conversations emerge and develop within larger conversations. The nature of these platforms, then, allows for people to employ collective sensemaking. Research has shown that in times of crisis or political upheaval, people often turn to social media platforms like Twitter and Facebook to search for information in the hopes of reducing uncertainty and making retrospective sense of what occurred, especially if information surrounding the phenomenon is low (Stieglitz et al., 2017). Further, social media has been shown to support sensemaking in the context of people interacting with and accepting novel technologies like blockchain (Amadoru et al., 2019). It seems appropriate, then, that sensemaking in these forms (attempting to understand the technology, interacting with others for more information) may play a role in the way people attempt to understand novel media forms like non-malicious deepfake technology

2.3.2 Contextual Integrity

Contextual integrity, a framework used to examine and determine privacy violations, sees privacy as rooted in contextual norms of appropriateness and norms of distribution (or flow) (Nissenbaum, 2004). As Nissenbaum (2004) explains:

> It is crucial to know the context-who is gathering the information, who is analyzing it, who is disseminating it and to whom, the nature of the information relationships among the various parties, and even larger institutional and social circumstances. (p. 154-55)

In a legal and ethical sense, this framing of information with privacy provides a clear metric by which privacy violations can be examined and determined; the amount of people's information being collected and shared (out-of-context) with and without their consent (distribution), and the *type* of information that can be shared within a specific context (appropriateness) determine if and how there has been a transgression (Nissenbaum, 2004). Under contextual integrity, it would be appropriate to share sensitive medical information with a doctor, but not with a manager. Following that same line of thinking, a doctor could distribute medical information to fellow medical staff where appropriate; violation of flow would occur if this information were shared with an advertiser. Contextual integrity has been used to examine privacy concerns in emerging technology, like RFID technology and data mining (Nissenbaum, 2004), and IoT technology (Jia et al., 2017).

Malicious deepfake technology, especially nonconsensual pornography, may violate norms of appropriateness and distribution by using mass-collected images, removed from context, inserted into a different context, and presented as real. And, although non-malicious deepfakes are developed without intent to harm, questions of ethics and privacy remain, particularly if those featured within a deepfake video did not consent to the sharing and manipulation of their image.

2.3.3 Audience Theory

Webster's (1998) Audience Theory suggests that relationships exist between mass media and the public (also called *the audience*). These *audiences* can be categorized into three different types: *audience-as-mass*, which examines the relationship between the types of media large groups of people consume and the impact it has on the group, *audience-as-agent*, which focuses on the agency of people to select and consume media within their own social and cultural contexts, and *audience-as-outcome*, which views the relationship between people and media as people being acted-upon by media (Webster, 1998). Audience-as-outcome in particular sees audiences as particularly influenced by the media they consume (*e.g.* consuming violent media may change a person's perception that "the world is more violent") (Webster, 1998). These three forms of *audience* are not always mutually exclusive. Rather, Webster argues instead that these audiences can sometimes overlap in certain circumstances (Webster, 1998).

Active Audience Theory, which argues that people do not simply receive and accept media (and information), builds in part upon the encoding/decoding model presented originally by Hall (2000). In the encoding/decoding model, a message is encoded with a particular meaning, which is then received and decoded by the recipient; the meaning in the message may differ from what was originally sent (Hall, 2000). Encoding/Decoding also suggests that power dynamics play a part in how the message is encoded and decoded; audiences may accept the dominant narrative, negotiate meaning (while still accepting a portion of the dominant narrative), or reject the dominant meaning/understanding outright, constructing their own interpretations (Hall, 2000).

The rapid emergence and spread of novel forms of digital media (like non-malicious deepfakes and other synthetic media) have presented new challenges to Audience Theory and the Encoding/Decoding model, and the subsequent research built on these models. First, while social media platforms are considered a form of mass media, they also afford the creation of user-generated content (such as text) and discourse between users (Boyd & Ellison, 2007; Kunduru, 2018). This means that social media users are not simply consuming and sharing media and text produced by others; they also create their own. This also means what may be received in one way within this smaller context may not be perceived the same way in another space. Further, some newer digital media forms don't necessarily convey a direct message to be consumed and interpreted by the audience: indeed, the non-malicious deepfakes examined in this research are creative endeavors *without* a direct message (e.g., "non-malicious deepfakes are good", "creating non-malicious deepfakes is bad).

Still, revisiting both Audience Theory and the Decoding/Encoding model may prove useful not only in understanding how audiences respond and react to the non-malicious deepfake technology presented in this study but may also provide insight into how messaging (through framing and intent) changes between the content creator and the commenter viewing the content. While Encoding/Decoding has been used mostly to describe the relationship between mass media and the public, it might also explain how audiences react to non-malicious deepfake technology in ways creators and those sharing deepfake content did not intend. Further, the relationship between those creating and sharing non-malicious deepfakes and the people (audience) reacting to them suggests that applying a mixed-model approach, that blends aspects

of audience-as-agent and audience-as-outcome might explain how the framing of non-malicious deepfakes by creators as well as people's experiences and backgrounds influence reactions.

2.4 Gaps in the Research

Since its emergence in 2017, deepfake technology has become more widespread, and as indicated by prior research, has been used in mostly malicious ways. While there has been some research conducted on peoples' reactions to the use of non-malicious (creative) deepfake technology in museums and on videos sharing websites like YouTube, there has been less of a focus on people's reactions to these deepfakes on social media websites like Twitter and Instagram, where single users or organizations can influence and direct the conversation. Further, more research is needed that examines non-malicious deepfakes considering well-established theories developed in the fields of media, ethics, privacy, and computer-mediated communication. This research attempts to fill the gap by examining peoples' reactions to the non-malicious, creative use of deepfake technology while also positioning findings within prior theory and research on sensemaking, contextual integrity, and audience.

Chapter 3: Deeptomcruise

Special effects creator *df_creator* is a visual effects artist and content creator whose deepfake videos of well-known actor Tom Cruise have generated millions of views and tens of thousands of comments on Instagram and other social media platforms. Through interviews, *df_creator* maintains they are uninterested in using deepfake technology with malicious or harmful intent and instead hopes to illustrate how deepfake technology can be utilized to produce interesting content.

Figure 3

Screenshot From Deeptomcruise Video.

Yet reactions to *df_creator*'s videos have been mixed; many who come across these deepfakes remain genuinely tricked, unable to determine if these deepfakes show real events (no matter

how bizarre, absurd, or seemingly out of character). Others recognize and are amused by the deception and encourage future creation. These competing reactions point to the need for further investigation: What do these reactions illustrate about the nature of deepfakes on social media, and what discoveries can be made about the way people attempt to make sense of this novel technology?

To better answer these questions concerning RQ1 (how creators characterize their development and use of non-malicious deepfake technology) and RQ2 (how people react to non-malicious deepfakes), this research follows a case study approach, focusing specifically on posts made by *df_creator* on Instagram and news articles surrounding their work. Data examined came from two main sources: *df_creator*'s publicly available and widely publicized Instagram and news articles curated by the Google search engine's advanced search results.

3.1 Data Collection

Data Overview.

In the first phase of data collection, I focused on *df_creator*'s posts on the social media platform Instagram. Although *df_creator* posts their deepfake creations on multiple social media platforms (including TikTok, YouTube, and Twitter), Instagram represented a good balance between access to the data (TikTok does not display comments unless a user creates a TikTok account) and the total amount of data available (Twitter and YouTube have fewer comments and replies than the same content generates when posted on Instagram). To be included in the study, posts needed to meet a set of requirements, including:

1. **An intentional focus on *deeptomcruise***. Each post needed to include *deeptomcruise* or *deepfake* in the title, post description, or in a top reply by *df_creator*, and the main content of the deepfake video needed to feature *deeptomcruise* specifically.
2. **Period.** The dates of each post needed to be between January 6, 2021 (the date of the first Instagram post using *deeptomcruise*) and October 31, 2021 (the end of data collection).
3. **English language**. Replies to each post needed to be in English, due to a lack of resources for accurate translations.

Instagram Data.

In December 2021, I downloaded posts directly from *df_creator*'s Instagram and saved each as .html files. Videos were individually downloaded from each post and saved in .mp4 format. Data saved but excluded from the analysis included individual user profile photos and other page files (i.e., .css and .js files.

Table 1

List of Deeptomcruise Videos

No.	Date	Description
1	*6-Jan-21	America deserves a real Top Gun in office A president that knows how the fly a fighter jet and knows how to make a perfect cocktail... #RunTomRun #Deepfake #VFX
2	7-Jan-21	Wanna know what it takes to make Tom run? Check out this VFX breakdown #RunTomRun #Deepfake #VFX
3	15-Jan-21	TomCruise2020 Concession Speech. Deepfake Breakdown. #deepfake #breakdown #tomcruise #tom2020 #artificalintelligence #vfxbreakdown
4	13-Mar-21	MY IMPRESSION... #deeptomcruise
5	26-Mar-21	Still got it.. #deeptomcruise
6	3-May-21	A #tiptok from our very own #deeptomcruise.
7	10-May-21	Deeptom's got a sweet spot. #deeptomcruise @deeptomcruise_ai
8	23-May-21	@deeptomcruise_ai keeps his hands clean. #deeptomcruise
9	11-Jun-21	Tom's is off the charts... #deeptomcruise
10	14-Jun-21	My man @deeptomcruise_ai got his stache on point.
11	23-Aug-21	Look who's back! #deeptomcruise
12	27-Aug-21	#Deeptomcruise got some magic coming up for you..
13	29-Aug-21	#deeptomcruise jamming! Making real music again!
14	2-Sep-21	#Deeptomcruise on the road!
15	9-Sep-21	#Deeptomcruise rented the entire Grand Canal. What a guy.
16	15-Sep-21	Never too old to learn! #Deeptomcruise
17	21-Sep-21	Dance-off ! #Deeptomcruise
18	1-Oct-21	#deeptomcruise taking Spanish lessons. Did he do well?
19	7-Oct-21	Deeptomcruise x Justin Bieber Who did it better?
20	12-Oct-21	#deeptomcruise likes a well fed crew!!!
21	29-Oct-21	Here's how #deeptomcruise suits up!

Note: *This post (and its 41 comments) was erroneously excluded from the analysis. Emoji were removed from the title where needed.

To clean and prepare the .html files for analysis, I opened each .html file in a text editor. Iteratively, I used a series of regex to pattern match usernames and comments within the .html file.

Table 2

Steps to Collect Data from HTML Files

Step	Regex / Manual	Description
1	<span class="Jv7Aj.*?href="https://www.instagram.com\/.*?<div	isolate username and text
2	href=.*</	get everything at instagram url
3	tabindex=.*	get everything from tabindex that holds username/text
4	manually remove tabindex="0">	
5	manually remove < /	
6	manually remove span> </h3>	Find and replace with
7	manually remove < /	blank space
8	manually remove <a class="notranslate" href="https://www.instagram.com/	replace with 'response:'
9	manually remove /"	Find and replace with
10	manually remove a>	blank space
11	manually remove <a class=" xil3i" href="https://www.instagram.com/explore/tags/	replace with 'tag:'
12	remove "#"	Find and replace with blank space
13	copy username and text and paste into .txt file	
	Miscellaneous regex used as needed	**Description**
	^(.*?)	everything up to comma
	[\s\S]*$	everything after comma
14	(?<=tabindex="0">)(.*)(?=<\/a>)	get username/handle
	</a(>[\s\S]*)$	clean up for final usernames
	(?<=<span class="">)(.*)(?=<\/)	get text from user

Following these steps, I was able to capture usernames and comments (as well as additional text that would be deleted where necessary – see miscellaneous regex in Table 3.1.2). This text was copied from the .html file and saved to both a .txt file and an Excel file named for the date of the original Instagram post. At the time of data collection, there were a total of 23 posts on Instagram that featured *deeptomcruise*, 20 of which were included for data collection based on the requirements listed above.

A total of 8,758 comments were collected over 20 posts. Unique usernames were not counted and were used primarily to provide context for conversations and threads within a particular post. The data also included many emoji responses. It was often difficult to determine what emojis meant without supporting text. As such, emoji-only responses were excluded from the analysis. However, emoji accompanied by text was included.

Google Advanced Search.

In place of a formal interview with *df_creator*, I decided instead to examine interviews *df_creator* took part in with journalists and reporters. Using Google's advanced search feature, I searched the keywords *deeptomcruise* and *df_creator* filtering for search results only in English. I also set the date of the search to between 1/6/2021 (the date of the first *deeptomcruise* post on *df_creator*'s Instagram) and 10/31/2021 (the arbitrary date used to scope the research). In total, 6 articles matched these criteria. Using NVivo's NCapture feature, I converted each web page to a downloadable .pdf document and imported the files to NVivo.

Table 3

List of Deeptomcruise Articles.

No.	Article Title
1	Here's How Those Surreal Tom Cruise Deepfake Videos Were Made
2	'I don't want to upset people'~ Tom Cruise deepfake creator speaks out ~ TikTok ~ The Guardian
3	The Tom Cruise deepfake that set off 'terror' in the heart of Washington DC - ABC News
4	Those viral Tom Cruise deepfake videos were created by this man ~ Fortune
5	TikTok Tom Cruise deepfake creator~ public shouldn't worry about 'one-click fakes' - The Verge
6	Tom Cruise deepfake creator says the technology should be regulated ~ Fortune

Note. Article titles were generated from the NCapture process.

3.2 Data Analysis

Initial Approach.

To gain a better understanding of the data, I first focused on a subset of the data (called "pilot data"). I randomly selected 5 posts, then imported the comments into NVivo, a qualitative research analysis software. For each file, I used NVivo functionality to generate word clouds for the top 20 words longer than 3 characters, including stemming ("looks", "looking") and synonyms. Instagram handles and usernames, (@username), common English stop-words ("it", "it's"), and other miscellaneous text characters (i.e., text meant to generate emoji) were excluded.

Figure 4

Deeptomcruise Word Cloud

Note. This word cloud contains the top 50 words (exact match) from all Instagram data included for analysis.

I also generated word count lists as a separate means to gain a quick and functional overview of the data. The word count lists and word clouds were not analyzed or included in the findings.

Coding.

After surveying the data, I followed an inductive approach grounded in thematic analysis practices common to qualitative research (Braun & Clarke, 2006). I selected an initial random file, then coded in a descriptive style to capture the thoughts behind each reply, ensuring codes were specific and accurate to the way and intent with which people communicated. Initial codes included (but were not limited to): *who is that, looks real, confusion, conspiracy, deepfake tech, deceived, conflicted, critique, praise, determining truth, target thoughts,* and *reflection.* These initial codes included instances of people "making sense" of deepfakes (determining truth), reacting in an affective way (feeling confused), promoting communication between the creator

of the deepfake and commenter (via critique), and considering the ethical implications of the technology (via reflection).

With these initial themes in mind, I selected a new sample: for each Instagram post that had more than 100 replies, I randomized the comments and included 25% of this new randomized sample in the analysis. For posts that had less than 100 replies, all comments were included in the analysis. In total, 3,869 comments were coded and analyzed. Data from all 6 articles were included for analysis. During this stage, data saturation was reached; I saw repeating patterns in reactions and responses from commenters, and creating an additional sample to analyze data would have yielded redundant results. The remaining data from this sample was coded to already developed codes.

To analyze the data, I followed a mixed deductive and inductive approach, first coding the data to codes generated from the first round of coding. This second round was also more targeted: Guided by the themes from my first codes and my theoretical framework presented in **Section 2.3,** I specifically looked for instances of sensemaking strategies, ethical considerations of deepfake technology use (including consent and privacy), and interactions among the audience. I developed new codes as needed, and wrote memos and notes as I coded, focusing on similarities between the codes. I then grouped these as subcodes beneath a more descriptive and focused main code that tied back to my initial thoughts on sensemaking, ethics and contextual integrity, audience interactions, and affective reactions to deepfakes/deepfake technology. As more distinct patterns emerged, I also examined and coded comments within their specific contexts; that is, I examined comments as part of micro-conversations (i.e., *replies* to other commenters). Finally, I

grouped these main codes beneath broader themes directly tied to R1 (creator characterizations of non-malicious deepfake technology use) and R2 (people's reactions to non-malicious deepfake technology). These initial themes included: *relate deepfake to other media (sensemaking), attempt to determine or report truth (sensemaking/information seeking)*, *enjoying the deepfake technology for what it is (audience)* and *attempt to understand the ethics of the technology (ethics and contextual integrity)*.

3.3 Findings

Commentator reactions to non-malicious deepfakes varied, and fell into distinct categories: *emotion in responses*, *consulting, critiquing, correcting, focusing on the deepfake target*, and *working through the ethics of deepfake technology*. Creator characterizations of their deepfake technology use focused mostly centered on defining their technology use in terms of exploring technical and creative possibilities.

3.3.1 How people react to deepfake technology

Emotion in Responses

Many replies and comments centered on how the deepfake videos made them *feel*. These reactions were often short, often just a sentence, a few words, or a string of emojis. Reactions ranged from positive ("Too good, love these!"), to negative ("is there a hate button?") to decidedly mixed ("This is cool but also... incredibly frightening"). In the case of positive reactions, the knowledge that the videos they viewed were deepfakes did not appear to have much of an influence. Many people recognized the videos were indeed deepfakes, but still expressed positive emotions:

[commenter] I **loveeee**
[commenter] I **love** these so much
[commenter] So candid! I **enjoyed** this.

None of the videos were intended to be scary or offensive. Each video featured *deeptomcruise* engaging in normal activities, such as playing the guitar, doing a backflip, or enjoying a meal. Still, many comments expressed how *terrifying* or *creepy* the videos were:

[commenter] That is creepy af. These deep fakes….
[commenter] You're freaking me out.
[commenter] This is terrifying

Consulting, Critiquing, Correcting

Studies examining reactions to and interactions with deepfakes and deepfake technology have shown that people are not only interested in determining *what* is real but also in developing strategies and methods (what Bode (2021) calls 'critiquing' and 'correcting') to ensure that others who come across these deepfakes understand and can *judge for themselves* what is real. This study found similar results: many who reacted to the *deeptomcruise* videos used a set of consulting ("sensemaking") strategies to determine if the videos they watched depicted real events, while others familiar with deepfake technology, face-swapping, or similar apps used critiquing methods to point out when the video appeared to be a deepfake. These posters also corrected those who were still unsure or believed the deepfake videos to be real or depict authentic events.

Consulting

In the following examples, each person's confusion seems to stem from other comments in the thread that indicated the video does not feature the actor and instead utilizes some type of impersonation. Posters used consulting techniques (questioning, indicating confusion) to ask for help determining the truth:

> **[commenter**] Wait!!!! In the comments people says that he is NOT Tom Cruise??????? I don't understand
> **[commenter]** Wait…is this Tom Cruise or a look alike?
> **[commenter]** Is this the real Tom or not?? I really can't tell

Others mirrored these reactions, expressing a mix of genuine befuddlement while appealing generally and directly to those in the thread to help them make sense of what they saw:

> **[commenter]** So, this is actually NOT Tom Cruise? Help me understand???
> **[commenter]** please explain why this is terrifying. I don't know what I'm watching. Is this not Tom Cruise? If not who is it? So confused.

Some seemed aware of the existence of deepfake technology or similar "face swap" applications developed to "mimic" the faces and voices of others, but like those who noticed differences between the deepfake and the real person, they ultimately remained undecided:

> **[commenter]** Wait I don't get it! Is this a deepfake? Or is this a guy who looks like tom cruise???
> **[commenter]** is this voice deep faked or is the guy a good impressionist??

> **[commenter]** love it I heard there is a filter someone is using to "be Tom cruise" or Mimick... idk what you would call it. Wonder if this video is real?

Critiquing

Many comments focused on critiquing the technical aspects of the video itself. Some used critique to distinguish these deepfakes from "real" videos that had not been altered with deepfake technology or special effects. These comments pointed out breakdowns in technique that made the video seem not quite right:

> **[commenter]** Smooth but **lighting or color seems off.** Still awesome though
> **[commenter] Mouth doesn't quite line up**. But soon...
> **[commenter] Ok the shadow under the chin** isn't done 100 percent. I was close to fooled.

These comments served to point out *areas of improvement* and to provide feedback that could perhaps be integrated into future videos. People also wanted to share instances where they were impressed by the techniques used:

> **[commenter]** Holy crap lol! **That's insane how well the effects are blended.** I had to look up Deepfake that some of y'all mentioned and was surprised!
> **[commenter]** This is a really good deep fake **not as flat faced as they normally are also the lighting is good.**

These comments were encouraging, mixing critique with a genuine appreciation of the technical skills used to create each deepfake video. Comments in this category also often crossed over into direct praise of *df_creator*:

> **[commenter]** **Love how you make a point of doing things that a deepfake would usually struggle with.** That's what really sells the effect. It's like the magician's assistant showing you an empty box before the magician climbs in and has swords shoved through.
> **[commenter]** Wow even the side angle of the Face Looks so flawless and convincing! **Great work I'd say that's the best deepfake I've ever seen**

Correcting

Replies to each video contained a fair mix of people at varying levels of understanding of deepfake technology and use. This created a space where those who recognized the deception were able to correct those who did not. These correction techniques varied in tone and length, but all emphasized the "wrongness" of those who assumed the video depicted authentic events:

> **[commenter]** you really haven't heard of it? Lol look it up**. (informing)** That's why it says #deeptomcruise as in deepfake. It's also why he has the voice of an entirely different person lol…
> **[commenter]** No it's not. All these videos of him coming out lately are DEEP FAKE Video with his face superimposed on this guy which is why his voice is always way off.

Like those who critiqued the technical aspects of the video, those who corrected often compared the actor Tom Cruise with the deepfake creation to point out breakdowns:

> **[commenter]** Actually **I did a snapshot of this guy and compared to Tom**. This guy is NOT Tom.

In this next example, the poster (incorrectly) concludes that the video depicts the real Tom Cruise, basing this on their direct comparison between the deepfake/impersonator and the real

person:

> **[commenter]** This IS Tom. **They have different noses.** You really have to look at the shape of the nose cartilage.

In the example below, [commenter] uses informing and comparison to correct others, even pointing out the #deeptomcruise hashtag used to signify the video is a deepfake:

> **[commenter]** ***(informing)*** Just research it it's no secret at all. Google something like deepfake faceswap or something like that…There's many videos just like this in which they superimpose someone's face onto someone else seamlessly in real time in a video…***(correcting)*** To reiterate this is not Tom Cruise. ***(comparing)***1. Tom Cruise has a completely different voice…***(informing)*** The caption of the video uses the hashtag #deeptomcruise to signify that they're doing this very thing ***(providing resources)*** If you don't believe me Google it or simply click on #deeptomcruise hashtag to see more videos just like this…*

Micro-conversations also emerged within the replies as those responding worked together to reach a consensus:

> **[commenter1]** ***(consulting)*** Please explain to me what's happening?? Is this the real Tom cruise or some bloke that looks just like him??? His voice isn't Tom Cruise's.
> **[commenter2]** ***(correcting)*** it's a deep fake
> **[commenter1]** So he doesn't actually look like that?
> **[commenter2]** ***(informing)*** nope there's another dude there and talking and they use computer software to overlay Cruises face on top of the performance. The computer also downloads tons of reference of Tom to aid it. ***(offering resources)*** Just YouTube 'deep

> fake' the tech is quite cool and also scary at how good it's getting and even surpassing CGI attempts at faces
> **[commenter1]** ***(indexing/confirming)*** Ok thanks for the info I had no idea this was a thing. I can't believe how real it looks

In the example above, *commenter 1* noticed a difference between the "voice" in the deepfake video, and the voice of the real person, but was confused about how this could be possible. *commenter 2* helped *commenter 1* make sense of what they saw by explaining the concept of "deep fake", while also providing additional resources for *commenter 1* to do further investigation on their own. With this new information, *commenter 1* confirmed their understanding that what they saw was indeed a deepfake. This type of back-and-forth pattern (consult, inform, correct, confirm/index) representing a dominant sensemaking strategy will be explored in Chapter 5.

A Focus on the Target

Many of those reacting to the deepfake videos on *df_creator*'s page included references to the actor's (perceived) personal and professional life and background, and in some cases, wrote comments as if they were directly interacting with him:

> **[commenter]** Hypocrite Tom Cruz! **Abusing your own staff who shows up at the job for tiny fraction of what you make!** No respect for **you** Tom Cruz and **apologize to those whom you abused out of your inexcusable hates and angers**!

This comment coincides with an event during which the actor admonished staff for not following Covid-19 protocols on a movie set (Limbong, 2020). Posters also directly petitioned the actor in

a bid to influence their real-life choices and decisions:

> **[commenter]** Damn it you are so beautiful on the eye yet you live such a restricted life of false belief and surface gain it's such a pity because you could be a hundred of you and show the world such goodness with your success and influence

While it is unclear whether those in these examples recognized the video as a deepfake (and they were not sending messages directly to the actor) they used their ability to reply as an opportunity to voice their frustrations.

Working through the Ethics of Deepfake Technology Use

Many individuals wondered about the legality of deepfake use, with some expressing strong moral objections against the use of the technology. In some cases, people advocated for more direct action against deepfake technology, either through banning the technology outright or punishing individuals who used the technology without the consent of those featured in the videos:

> **[commenter]** This is scary and it **should be illegal.**
> **[commenter]** Deepfake **should be illegal.**
> **[commenter]** I can't believe Deep Faking as another living person is legal…
> **(punishing) [commenter]** This is so dangerous what you are doing here. **Hope you get caught and jailed.**

Others drew a comparison to accounts being shut down or banned. Although the meaning of "truth advocates" remains unclear, [commenter] advanced the idea that deepfakes should be

considered a threat:

> **(banning) [commenter]** But why aren't these accounts being shut down the way truth advocates are? Deepfake is such a huge threat to humanity

By contrast, the data showed no instances of people directly advocating for deepfake technology to be *legal*. What's more, when discussing questions of morality and ethics, the data did not show instances of people affirming the morality of deepfake technology; replies directly centering deepfakes as moral did not exist. Rarely, replies focused on consent, specifically whether the actor agreed to or knew deepfakes featuring his image were being produced:

> **[commenter] Does Tom cruise know** about this? Lol
> **[commenter]** I just love these so much. Has Tom Cruise actually ever acknowledged them? **I wonder what he thinks!!!**
> **[commenter]** I love this but **can't help but wonder Tom's position on his deep fake persona. Is he OK with it** or watching in the background for opportunity to strike?

While these posters seemed interested in understanding the actor's perspective, they stopped short of classifying or identifying the creation of these deepfakes as nonconsensual or morally wrong. Still, many individuals expressed a general worry that deepfake technology itself might cause future harm:

> **[commenter]** Yeah this tech will never lead to anything bad happening…

In the case of one response, parallels between deepfake technology and other world-changing advances in tech seemed clear:

> **[commenter]** This AI shit is worst than the atomic bomb. It will mess up our world for real

In the examples below, one commenter appears shocked that others don't seem to realize the videos are deepfakes, with another worrying about what this lack of awareness or realization might mean:

> **[commenter]** The fact that people don't understand what this is and the technology behind it **is what really scares me**. EVEN with the caption...
> **[commenter]** Based on the comments **I'm worried** that more people don't know what deep fakes are…

These responses underscored a general feeling that although the technology is interesting and cool, it has the potential to cause harm.

3.3.2 Creator Characterizations of Deepfake Use

Despite rich conversation surrounding these deepfakes, there appeared to be little direct interaction between *df_creator* and those reacting to the deepfake videos on their Instagram page. In rare instances where *df_creator* appeared to interact with someone, the communication was short and appeared to only discuss technical details of the deepfake itself:

> **[*df_creator*]** Yeah you're totally right. I'm working on a solution for the eyes…I should have added motion blur. Will do that on the next projects. **Thank you for your honest feedback.**

Responding to another reply about the technical details of the video, *df_creator* again appears most interested in improving upon technique:

> **[*df_creator*]** Thanks a lot for the kind words. I've put a lot of post in this one to get certain angles to work… But as you said it's all about challenging the impossible.

In interviews, *df_creator*'s motivations and intent were less hidden. In particular, *df_creator* remained clear that the *intent* behind their deepfake technology use was to explore technical and creative possibilities (in their field of visual effects and filmmaking):

> **[*df_creator*] It gives you so much more creative possibilities**…If you have an actor who would be amazing for a role but doesn't physically fit the role…they can now play that role.*
> **[*df_creator*] My intention is clear**, and that is to make fun videos…**I have no intention to fool anyone or fool the system.**
> **[*df_creator*] I don't intend to use it in any way where I would upset people** – I just want to show them what's possible in a few years."

And, although it is unclear what *df_creator* means by "the system", what is certain is that they do not want to create deepfakes as a means to "fool" anyone beyond reasonable use of the technology. This is perhaps the reason why *df_creator* appended the word "deep" to their deeptomcruise creation and indicates through use of this word in other places (such as posts) that

they are not meant to be real and are deepfake creations.

Like those replying to the deepfakes, *df_creator* also seemed at least partially aware of malicious deepfake technology use, particularly as the technology becomes democratized and more widespread. On this, *df_creator* appeared to agree that deepfake content needed to be regulated in some form:

> **[*df_creator*]** I just strongly think **that there should be laws to help with the responsible use of AI and deepfakes**

3.4 Chapter Summary

In this chapter, I examined people's reactions to *deeptomcruise* on Instagram. In particular, the data showed that people engage in information-seeking and sensemaking behavior while trying to understand deepfake technology. My findings also show that people had affective (positive) reactions to *deeptomcruise*. I also found that some commenters raised ethical questions about the use of deepfake technology, particularly surrounding whether consent had been given to collect and use the photos, with others agreeing that the technology should be banned. And even though many commenters were excited and awestruck by the deepfake technology videos, there was no conversation surrounding deepfakes being made explicitly legal. By examining news articles, I also focused on *df_creator*'s motivation behind their use of the technology and creation of *deeptomcruise.* The research showed that *df_creator* saw their use of the technology as a means to demonstrate future possibilities in media and film.

Chapter 4: DeepNostalgia

Created by D-ID, an AI company that specializes in “creative reality”, DeepNostalgia is an ongoing project that emerged on Twitter in February 2021. Developed in collaboration with the genealogy website MyHeritage, the app allows people to upload still images to a central database, where they are transformed into videos. The process by which these videos are animated uses D-ID’s “Live Portrait” technology, which itself is a form of deepfake technology. Created primarily for those interested in transforming private family photos, the technology has also been used to animate well-known historical figures like Frederick Douglass.

Figure 5

Deepfake of Frederick Douglass.

Note. Screenshot from a deepfake video of Frederick Douglass created with DeepNostalgia app (Bruce, 2021).

The use of deepfake technology in this way raises key questions: What differences (if any) do people perceive between deepfakes they may have created of their relatives, and those of celebrities and historical figures? What factors may influence their reactions?

4.1 Data Collection

In the first phase of data collection, I examined a collection of Tweets surrounding the release of #DeepNostalgia and the Live Portrait technology. Twitter provides access to its stored data in two ways: through limited search that only looks for tweets within a specific timeframe (usually only within the past 7 days) and through full archive search which contains all tweets ever created. I pulled Tweets from the full archive via twarc2, a command-line tool written in Python used to retrieve JSON data from Twitter. To be included in the study, posts needed to meet a set of requirements:

1. **Use of the keywords [#] deepnostalgia and the phrase "My Heritage animation".** Not all tweets used *deepnostalgia* in any form or contained "My Heritage animation" within the tweet text, but these keywords were chosen to narrow the scope and provide focus to the research.
2. **Time period.** To capture this data, I pulled tweets and Twitter data from the period between 2/22/2021 and 3/22/2021 (the end of data collection). The start date coincided with the release of the Live Portrait technology through the MyHeritage website.
3. **Use of deepfake media within a conversation.** Tweets included in this study needed to have as the top post (start of the conversation) a deepfake video created by that poster.

This allowed me to collect data as part of specific “conversations” related to a specific deepfake video.

4. **English language**. Replies to each tweet needed to be in English, due to a lack of resources for accurate translations.

The table below shows the general form and operators included in the search query used to collect Tweets and Twitter data.

Table 4

Keywords and Operators in Twitter API Search

hashtag/phrase	account	attribute
#deepnostalgia #DeepNostalgia deepnostalgia “Deep Nostalgia" "deepnostalgia” My Heritage animation	conversation_id	has:media -is:retweet lang:en

To build the final query, I used the Twitter Advanced Search tool to search for relevant tweets defined in the previous section. In Jupyter Notebook, I wrote a simple script in Python, using the twarc2 command-line tool to pull Twitter data from the full archive. I then compared the two result sets, making sure Tweets collected from the Advanced Search generally matched those pulled using the script. I completed this step multiple times, comparing the Advanced Search results and refining the query until the Advanced Search results and full archive search aligned. In total, 10,628 tweets were collected between 02/22/2021 and 3/22/2021.

The code below shows the general form (including search query parameters and file name) for the search query used to collect data, where “start time” and “end time” represent the start date and end date, and “file_name” represents the name of the file used to store jsonl data pulled from the Twitter archive:

```
# search tweets
!twarc2 search "(#deepnostalgia OR #DeepNostalgia OR deepnostalgia OR ""Deep Nostalgia"" OR ""deepnostalgia"" My Heritage animation) -is:retweet lang:en" --start-time "start time" --end-time "end time" --archive --limit 100000 file_name
```

To select a sample of the data for analysis, I focused on the top 10 tweets by the number of replies (descending) for each target week. In total, this represented 40 of the most replied to tweets. I then used the *conversation_id* from each top tweet set to run subsequent Twitter API searches to find all other related tweets related to the *conversation_id* (what Twitter calls “conversations”). For these conversations, the end time date was set to 4/30/2021 to capture all replies. Following this process, a total of 2,134 tweets were included for analysis. Although usernames were collected, they were not included within the analysis and were used primarily to provide context for discussions happening within a particular conversation.

Articles.

During an initial survey of articles and reports surrounding DeepNostalgia and Live Portrait, I found that the vast majority simply referred to or directly quoted information publicly available on the creators’ websites. As such, to better understand how MyHeritage characterized #DeepNostalgia/Live Portrait technology and use, I decided to examine information from both

websites. Using Google Advanced Search, I pulled articles directly from the D-iD (d-id.com) and MyHeritage (myheritage.com) domains. I restricted my search to the period between 2/22/2021 and 3/22/2021, coinciding with the general release of the technology, as well as the creation dates of tweets examined in this study. In total, 4 web pages matched these criteria.

Table 5

List of D-ID and MyHeritage webpages.

Post	Page Title	Site
1	Deep Nostalgia™ is an Internet Sensation! - MyHeritage Blog	MyHeritage
2	D-ID's Live Portrait Product- Embedded in MyHeritage System ~ D-ID AI face Platform	D-ID
3	New~ Introducing Deep Nostalgia™ — Animate the Faces in Your Family Photos - MyHeritage Blog	MyHeritage
4	What is Deep Nostalgia™~ - MyHeritage Knowledge Base	MyHeritage

Note. Page titles were generated from the NCapture process.

Using NVivo's NCapture feature, I converted each web page to a downloadable .pdf document and imported the results to NVivo.

4.2 Data Analysis

Approach.

I first focused on an initial subset of the data. Next, I imported the comments as text files into NVivo. For each file, I used NVivo word analysis features to generate word clouds for the top 20

words longer than 3 characters, including stemming (“looks”, “looking”) and synonyms. Twitter usernames (@username), common English stop-words (“it”, “it’s”), and other miscellaneous text characters (i.e., text meant to generate emoji) were excluded, as these features were not a focus of this study. I also generated word count lists from the word cloud as a means to gain a quick and functional overview of the data. The word clouds were not analyzed directly, and were used to gain a broad overview of the data:

Figure 6

Twitter Word Cloud.

Note. This word cloud contains the top 50 words (exact match) from all Twitter data included for analysis.

Coding.

Like the coding process outlined in Chapter 3, I followed a mixed deductive- inductive approach grounded in thematic analysis practices. I selected an initial random file, then coded in a descriptive style to capture the thoughts behind each reply, ensuring codes were specific and

accurate to the way and intent with which commenters communicated. Tweets were analyzed within their specific batch (e.g., all 2/22/2021 tweets were coded within the same file), but each tweet was examined individually. Initial codes included: *comparison, conspiracy, how does it feel,* and *questioning technique.* I also took notes and memos, and reviewed relationships between my initial codes, making note of how these relationships related to previously constructed themes of sensemaking, ethics, audience, and affective reactions to deepfakes/deepfake technology. As more distinct patterns emerged, I also examined and coded tweets within their specific contexts; that is, I examined tweets as part of conversations (i.e., *replies* to other tweets). I then developed subcodes, and finally, grouped these subcodes beneath final themes with respect to RQ1 (creator characterizations of deepfake technology use) and RQ2 (people's reactions to non-malicious deepfakes). This next set of themes included: *relates deepfakes to popular media (sensemaking)*, *attempting to define current and future uses of technology (sensemaking)*, *attempting to understand the ethics and philosophy of deepfake use (ethics and contextual integrity)*, and *hope to tell stories that provide context about the target's life (audience).* During this stage, data saturation was reached; the remaining data from this sample was coded to already developed codes and subcodes.

4.3 Findings

4.3.1 Contextual Factors Influencing Reaction

Many of the reactions examined in this study were made in response to a main deepfake video that depicted the uploader's (sometimes deceased) family member. People overall appeared open and receptive to these videos. In many cases, commenters who had used the technology to

create deepfakes depicting their own family members also shared their thoughts. These replies to the main post often included small snippets and details about the lives of those depicted in the videos:

> **[commenter]** My third cousin three times removed - Brian Hatton. **Born in 1888 and died in 1916 - a casualty of WW1. A successful painter,** who would have even more renowned if it weren't for **his untimely death**.

> **[commenter]** It's rather bizarre I must admit. **I never knew my grandfather Warrant Officer George William Smith** and to see him almost resurrected **is quite amazing...**

This "storytelling" and "information-giving" about ancestors and family members happened regardless of whether the deepfake video featured an ancestor many times removed, or a close relative (such as a parent or grandparent):

> **[commenter]** I did one of my father yesterday. **It is so incredible to see him move, smile, blink again.** He died at 44, massive heart attack out of the blue, 1973. He worked and fished. Didn't talk much to us kids, **but I miss him.**

Like others, [commenter] did not give a direct reason for using the app. What's clear is that their reaction *(it's so incredible*) was influenced by seeing their (deceased) father "move, smile, blink again" by virtue of the technology. Others had similar reactions:

> **[commenter] It was so good** to see my grandparents and watch their expressions again **after all these years.** This app **really moved my mom, they were her parents**, so kind of heartwarming too.
>
> **[commenter]** Yes, it's unsettling but it did bring a tear to my eye after watching some of my grandfather's pictures brought to life**. Seeing a smile flash across his face made me miss him even more.**

Despite the "uncanny" nature of the video, they still were "moved" by seeing their grandfather animated (smiling) in this new way. The appeal of this technology and its ability to "reanimate" deceased relatives was not lost on others:

> **[commenter]** Omg! I would love to do this with relatives that have passed and **don't have any video footage.**
>
> **[commenter]** I think it's kind of cool **especially for the younger members of a family who never had the chance to meet their relatives**, I would have it done for a member of my own family

Conversely, some people seemed less receptive to the videos. Many also rejected the notion that a deepfake could accurately capture the nuances of themselves or their family members, and this realization appeared to influence their reaction:

> **[commenter]** It can emulate human movement but **it can't imitate the persons unique movements** so **it looks like a person I know and love being poorly operated by alien.** Not nostalgic. **Just creepy.**

In this case, the commenter decided that the movements of the deepfake image were too dissimilar to those of their family members. Others agreed:

> **[commenter]** There a strange lifelessness about them. It's as if they've been created at that point in their lives but don't know what to make of it.

> **[commenter] unless you are animating the face with face acting data from the person (now older) from the photo, they aren't gonna have the same mannerisms** so 100% creepy uncanny valley. Could be cool for museums but not for a picture of grandma on the wall

This distinction (perhaps this app is more suited for museums and historical figures) was mirrored in other responses:

> **[commenter]** It depends on the photo and also if you knew the person or not. The photos of the relatives I knew tended to be off because the way they smiled, moved their head or blinked **was not a mannerism that the relative actually had vs if you didn't know them it seemed better.**

> **[commenter] I saw Harriet Tubman this morning. She looked alive!** I don't think I can take seeing a photo of my mom, though. She passed away when I was 12.

The implication here is that closeness to those depicted in the image impacts receptivity: historical figures may be acceptable in a museum, but relatives and even ancestors are off-limits:

> [**commenter**] Ok, no... that's creepy. As if I don't already have enough dreams about dead relatives, **I don't need to see a beloved photo move.**

Historical Context

Like the storytelling and information giving mentioned in the previous section, some people responded to deepfakes depicting historical figures by providing additional explanatory details and background on the lives of those depicted in the video:

> **[commenter]** Nero is my favorite emperor. But just as Caligula he angered the Senate and they condemned both with a Damnatio Memoriea…And that worked well, because so many years later they are both envisioned as thru and thru bad.

> **[commenter]** Selvarajah Yogachandran was a leader of the Tamil Liberation Organisation. He was murdered in Welikada Prison on 25 July 1983. Kuttimani's tormentors "gouged out" his eyes, as he wanted to donate them so that they could one day see an independent Tamil state.

For some, the unnatural movements of the AI could be explained or even enhanced by historical context:

> **[commenter]** I assume at some point you'll be able to adjust expressions in a more realistic way. My interest is seeing certain photographed individuals in a way (a moving image) that post-dated their existence**. Smiles were almost non-existent in early photos-I was OK with Abe's sad one here.**

For this person, the historical context (smiles being non-existent) could explain Abraham

Lincoln's "sad" smile. And, as with deepfakes depicting family members, some realized that it might be impossible to capture the *essence* of these historical figures:

> **[commenter]** If one of these famous people was a gruff character, **we get none of the gruff.**

4.3.2 Reacting to and Interacting with Deepfakes

Polling the Audience

A common form of interaction between those who used the app to create a deepfake video and their followers (i.e., those who responded to the initial tweet with their own tweets) was represented by a direct audience "poll". This often took the form of the original poster prompting, "Is this creepy or cool?" Replies to this question were often very short, sometimes one word (creepy, cool, amazing, uncanny, both):

> (responding to "creepy or cool" prompt)
> **[commenter]** *raises hand wildly* CREEPY!!
> **[commenter]** It creepy af, datz my vote
> **[commenter]** A little bit of both.
> **[commenter]** I vote amazing!

Some responses were more in-depth, and provided background and reasoning behind their reaction:

> **[commenter] I didn't like it.** Too Uncanny Valley. And, I asked myself, does this add anything to my understanding of my ancestors? **For my part, it added nothing** (whereas colorization does add understanding)

Critiquing and Comparing

Like the findings in Case Study 1 that showed people were interested in critiquing deepfake techniques, people reacting to #DeepNostalgia deepfake videos also focused on the technical aspects of each video, pointing out or examining breakdowns in technique:

> **[commenter]** Neat tech. #DeepNostalgia did well with this photo of Balcha Aba Nefso. **However, the model produces noticeable perturbations around the eyes.**

> **[commenter]** It seems to struggle with images **if they're not symmetrical.** Why can't 'ears' just be reproduced like a mirror image rather than being blurred? The **colourisation function also dumps sudden purple blotches everywhere where there's a dark shade** too, sorry - otherwise it's great.

Like those reacting to *deeptomcruise*, people also tied the *#DeepNostalgia* creations to popular media. Harry Potter, which featured moving images in newspapers and talking portraits, remained a popular comparison:

> **[commenter]** Amazing. Straight out of **Harry Potter**.
> **[commenter]** I feel like I‘m in **Dumbledore‘s office** with all the former headmasters of Hogwarts in their frames. Amazing.
> **[commenter]** Wow! We have a wall of our gene pool. I showed this story to the hubs, saying, "what would that look like on that wall?” He replied, “It would look like a **Harry Potter** movie.”

Some comparisons were more dystopian:

> **[commenter]** This is so cool. I want to watch **Blade Runner** now
> **[commenter]** Now all of our deceased relatives can be dead eyed ghouls for all eternity. **Sometimes Black Mirror gets it right.**

Ethical Considerations in Use of Technology

Many people reacting to deepfakes had a general sense of the ethical concerns and dilemmas presented by the emergence of deepfake technology, with some debating how deepfake technology might create (future) harm at the societal or global level. Many tweets pointed to parallels with what they saw as current trends in misinformation and malicious use:

> **[commenter]** Technology is great But the **wicked intent worries me when in the wrong hands**. Today, lies are used **in radicalizing, brain-washing & fomenting riots, harm, etc.**

> **[commenter]** So, um, y'all can see how this potentially will negatively impact your own members' future job security, **as well as our collective national security**, yes?

Another poster took this idea a step further, focusing on the impact of the technology on truth itself:

> **[commenter]** Terrifying. We are reverting back to a time wherein the only communication that is genuinely believable **is face to face,** for better or worse. **Society's transition as it struggles to learn this fact, will be very dangerous.**

This general worry (that deepfake technology could have a very widespread negative impact on

society and truth) was mirrored in reactions that focused on more specific, personal harms deepfake technology might cause for individuals:

> **[commenter]** Kinda scary to think about. Imagine someone with the means to seamlessly create a video from your picture to frame or **implicate you in a crime**

> **[commenter]** It can also be misused mam…as much as it could be a boon, it has only proved to be more of a curse. **Imagine the way someone could spoil someone' life with that amount of technological advantage.**

In both instances, the tweets display a general unease with the understanding that deepfake technology could be used to ruin the lives of others.

Not all commenters saw deepfake technology as overwhelmingly harmful, with some focusing on the potential benefits of deepfake technology use and its potential to impact others, particularly for those interested in interacting with the memory of family members who have passed away:

> **[commenter]** Imagine a mother **who's lost her young child who can now talk to her little girl** on her iPhone like Siri or Alexa but just **more real human-like conversation**

The potential benefit implied here is that a parent would be able to "interact" with their deceased child in a way that seems more "human" than the type of interactions one might have with a virtual personal assistant (like Siri or Alexa). And, although the use of the word "weirdness" implies a slight distaste for the idea, this poster seems to agree with the sentiment that deepfake

technology could provide a more “human” experience when interacting with deceased family members:

> **[commenter]** This is going to get weirder. Imagine your whatsapp history added to your picture or avatar. When you're no longer here, people can still talk to you via chatbots and voice assistants. **My digital me might be having dinner with my great grandkids one day.**

Like those considering societal and future harm, one poster saw envisioned people becoming stuck in a “rabbit hole”, unable to tear themselves away from deepfakes being used to help them overcome some trauma:

> **[commenter]** This is **VERY dangerous though**. it can be a great cathartic tool for therapists for people suffering trauma but some may not want to leave the rabbit hole and stay there forever. **This is the point where laws could be created to treat machines as living organisms**

This idea, that deepfakes can change expectations of what truth is, will be explored in Chapter 5.

4.3.3 Creator Characterizations of #DeepNostalgia

Characterizations and Intent

Through posts on their websites, the creators behind the Live Portrait technology and *DeepNostalgia* app were clear about how they wanted their product to be received and focused specifically on their intent behind developing this technology. Both wanted people to be able to generate content about their ancestors that would provoke strong emotional reactions from those who used and viewed the videos:

> **[D-ID]** In a bid to get their users closer to their family history, MyHeritage…wanted **to go beyond just still images.** They wanted to create **a real emotional resonance with the people in these old photos.**
>
> **[MyHeritage]** "You get a **'wow moment'** when you see a treasured family photo **come to life** with deep nostalgia. Seeing our beloved ancestors' faces **come to life** in a video simulation lets us imagine how they might have been in reality and provides a **profound new way of connecting to our family history.**"

In the examples above, there seems to be an understanding that these deepfake videos would produce emotional reactions *beyond* what might be experienced with a still photo; this expectation matches what I found in people's reactions to and conversations surrounding the deepfake videos generated using the app. Still, the creators also realized that the app could be perceived as unnerving:

> **[MyHeritage]** While many love the feature and consider it magical, others find it **uncanny** and are **uncomfortable** with the results.

Again, the creators seemed to anticipate or at least acknowledge that the emotional reactions to the technology would not always be positive.

Interestingly, the app and tech creators also appeared surprised by the uses and reception of their technology, both in its general impact on people, as well as in the ways people found uses of the technology beyond reanimating images of family members:

> **[D-ID]** When we developed this technology… we never imagened **the amazing impact it will have on people amazing results**
> **[D-ID]** We're the company behind the technology on this project, and **while we didn't intend it for statues we're glad it works well on them too**!

Perhaps recognizing that people using the product would want assurances that their data (personal photos of themselves and family members) would be protected, D-ID and MyHeritage did attempt to provide some guidance on how this data would be ethically handled.

> **[D-ID]** Working really closely with the tech team, we were able **to embed our solution into the heart of their system** – an on-premise solution which would ensure user privacy, which is a core to D-ID's values.

MyHeritage directly acknowledged its "responsibility" in making sure people who used the technology could distinguish between their unaltered photos and videos, and those generated by the deepfake technology app.

> **[MyHeritage]** We believe it is our **ethical responsibility to** make sure that people **see the difference** between simulated videos created using deep learning and original photos or videos.

Interactions with those using the app

Although rare, the creators behind the Live Portrait technology (D-ID) and DeepNostalgia (MyHeritage) app did sometimes interact with those who had used the technology to create their own deepfake videos. These creators also interacted with those responding to others' use of the technology. These responses mostly praised people who used the app to create deepfake videos,

and hinted at expanded feature development:

> **[MyHeritage]** We are happy that you are enjoying our technology! And it's only **the beginning**
> **[MyHeritage]** Nice! That's a **great use** of our technology! And it's only **the beginning**

4.4 Chapter Summary

In this chapter, I discussed people's reactions to the *#DeepNostalgia* hashtag on Twitter. I found that people engaged in storytelling and information-giving when discussing deepfakes featuring family members and historical figures. Further, the research showed that commenters were curious about the ethical and legal landscape not only of non-malicious deepfakes but of deepfakes in general, raising concerns about potential misuse and future harm. By examining blog posts and articles surrounding the #DeepNostalgia app and LivePortrait technology used to develop the app, I found that both MyHeritage and D-ID were interested in generating "emotional" responses from people while they used the app. And although they also believed they have a role in promoting the ethical use of the software, I found that when they did engage on Twitter with people using the app, the interaction centered mostly on promotion.

Chapter 5: Discussion

In this chapter, I examine sensemaking theory and the consulting strategies people employed while attempting to understand the appearance of these non-malicious deepfakes on Twitter and Instagram. Next, I explore the ethical dilemma posed by non-malicious deepfakes, discuss the ways people think about privacy, consent, and ethical use surrounding deepfake use and determine whether contextual integrity can be extended to the non-malicious deepfake context. I then re-visit Audience Theory and Encoding/Decoding considering the findings presented in Chapters 3 and 4 and evaluate how relevant the models may be within this and subsequent research. Finally, I present design implications inherent in the development of the technologies that make non-malicious deepfake creation possible and investigate how other frameworks and strategies that have been developed and used by creators and researchers in related fields might also be beneficial to use.

5.1 Sensemaking

When confronted with the *deeptomcruise* videos described in section 3.3.1 many people entered a state of uncertainty and confusion, prompting a need for some *type* of sensemaking. Despite the fact these deepfake videos were not threatening, did not present a crisis, were not related to an organizational challenge or technical task, and weren't developed in response to a global event (like COVID-19), the lack of information surrounding them nonetheless set the stage for information-seeking behavior. This information gap seemed in part exacerbated by the absence of a clear information leader or a central location where information could be found and examined; other than the use of the *deeptomcruise* tag, *df_creator* did not attempt to fully explain

the videos within the posts themselves, or via replies. Strategies people developed to make sense of the deepfakes mirrored those people employ on social media in crisis contexts: people reached out to others through replies or through micro-conversations to ask for more information that would reduce or eliminate their unsureness. As in crisis contexts, commenters were able to jumpstart sensemaking of others by pointing to specific ways they could learn more about the technology being used *("If you don't believe me Google it or simply click on #deeptomcruise hashtag to see more videos just like this"*). Through this consulting/sensemaking, people were not only able to arrive at the truth ("*he doesn't actually look like that?")* but also develop conclusions about the technology itself ("I ***can't believe*** how real it looks ").

Given the nature of the deepfake videos created from the DeepNostalgia app (people appeared aware that they were deepfake videos, and they often featured family members and well-known historical figures), it makes sense that sensemaking for information-seeking or task-completion was not as prevalent. As one commenter put it: *"If you don't pretend it's real and tell people it's not real, then it's really not that complex or controversial.*" Instead, people attempted to contextualize these deepfakes within their own experiences, comparing them to other media (like the moving portraits from the Harry Potter films or Black Mirror episodes), or telling stories about their family members featured in each video.

The role social media plays as the medium through which information from traditional media flows what (D'heer and Verdegem, (2015) call the "backchannel") is in its ability to magnify and move conversations from the physical to the digital (D'heer & Verdegem, 2015; Osborne-Gowey, 2014). Using hashtags (#) and mentions (@), people have also been able to direct

their opinions, from reactions to television shows to attitudes about upcoming elections. Hashtags like #BlackLivesMatter, #ows, and #ThisIsACoup have also proved useful for promoting conversation about social and political movements, creating affective "networked publics, a mobilized audience, identified, and potentially disconnected through expressions of sentiment" (Papacharissi, 2016). Networked publics are powerful enough to develop feedback loops of continuous engagement, but also ultimately tenuous, able to disband as easily as they were created (Papacharissi, 2016). Networked publics often employ sensemaking strategies to align their understanding (and influence the storytelling) of these collective events (Papacharissi, 2016).

"Action", as part of a networked public, collective or otherwise, was notably absent from conversations surrounding #DeepNostalgia and #*deeptomcruise*, perhaps in part because these hashtags are about the media itself, rather than being generated *in response to* some external event or scenario. Further, these non-malicious deepfakes were designed to be mostly unobtrusive and non-controversial, and while hashtags were used to generate and encourage digital conversation, the need to mobilize around them related to these deepfakes was ultimately unnecessary. As non-malicious deepfake use changes, however, mobilization and "action" may become more prevalent.

5.2 Contextual Integrity and Harm

As with malicious deepfakes, non-malicious deepfakes appear to violate norms of appropriateness and flow. In the case of *deeptomcruise*, vast amounts of data were collected to craft an entire deepfake persona around a real person, which was then used to depict that person

doing and saying things he never said or did. Although the *deeptomcruise* videos were generated using publicly available photos, those images were also removed from their original contexts. Further, despite people's genuine attempts to learn more about Tom Cruise's reactions to or involvement in the creation of these videos as illustrated in **Section 3.3.1**, it was never made explicitly clear on Instagram (through comments or replies by *df_creator*) what consent, if any, was given. In the case of the DeepNostalgia app, people were able to upload photos of living and deceased family members, ancestors, and others to the MyHeritage server, exchanging the (sometimes post-mortem) privacy of their relatives for the ability to interact with and use novel technology. Unlike public figures like Salvador Dali, who believed his image belonged to the Kingdom of Spain (even after death) (Mihailova, 2021), most people featured in deepfake videos were deceased, private individuals who did not appear to provide any sort of post-death consent explaining how their images could be used on social media or with novel technology. There is also a lingering argument that when people post photos to social media, those photos become part of the "public", meant to be consumed by others. However, the use of deepfake technology to create these videos removes control from those posting and sharing photos to those using the photos in ways not originally intended (Haasjes, 2018). Further, people often engage on social media as part of a networked public, extending privacy from the individual only to specific groups within specific contexts (Marwick & Boyd, 2014). As one commenter remarked: "*Creepy…plus what about the rights to the images*?"

Beyond privacy violations, malicious deepfakes (those designed to spread disinformation, or those created as revenge/non-consensual pornography) have been shown to have harmful effects, not only on those featured in the deepfakes but also on the concept of "truth" (Rini, 2020). And

like unintentionally spread misinformation, deepfake technology developed without malicious intent (like those generated by *df_creator* and the DeepNostalgia app) may result in increased harm. Consider, for example, a *deeptomcruise* video, in which the "fake" Tom Cruise announces he is ending his run for United States president in the 2020 elections. While it is unknown what impact this deepfake video had on people's perceptions of the 2020 election, what is known is that deepfakes may decrease people's trust in (news) media overall (Vaccari & Chadwick, 2020). That deepfakes, even non-malicious ones developed without intent to harm, hold the power to erode what people think of as truth, was not lost on many of those reacting to #deeptomcruise deepfakes; even user-generated content (like the deepfake videos created via the DeepNostalgia app) raised concerns. As one person speculated: *"Society's transition as it struggles to learn this fact, will be very dangerous."*

Finally, an interesting phenomenon occurred with those reacting to *deeptomcruise*: many commenters began interacting with the deepfake as if it were the real person. Although many of the interactions were positive and encouraging, others were decidedly negative or harsh ("*No respect for you Tom Cruz*"). While this research did not focus specifically on parasocial relationships – a phenomenon in which people believe they are in a relationship or have a connection with a media persona despite no such relationship or connection (Hartman, 2016) – one could imagine a scenario in which this type of relationship might develop between a person and a deepfake persona created from the images of a non-consenting or unaware target. If lines become blurred between the real person and the deepfake, this could have troubling implications for people who might not have the same legal, social, or financial protections that a famous actor might.

5.3 Audience

For both *#DeepNostalgia* and *deeptomcruise*, *df_creator*, MyHeritage and D-ID appeared interested in normalizing the use of non-malicious deepfake technology, centering the videos and apps within conventional use. For *df_creator*, the use of deepfake technology could be used to entertain and demonstrate novel technical capabilities. For the creators of the DeepNostalgia app, deepfake technology could connect people to their ancestors and family in a new and interesting way. This "normalizing" of the technology was often rejected: for many commenters, the videos were confusing, upsetting, and warranted discussions about current and potential future harm, as illustrated in Section 3.3.1. And, although D-ID and MyHeritage maintained that the app, while unnerving to some, ultimately provided "a profound new way of connecting to our family history", others rejected the narrative that even non-malicious deepfake technology could be normalized in this way; commenters focused on the ethical implications of deepfake technology use, connecting it to broader issues like the spread of misinformation and disinformation on social media. Commenters did not simply consume these deepfakes at face value as presented by *df_creator*, MyHeritage, and D-ID, and instead decoded them within the context of their own (unstated) experiences and understanding. Conversely, the positive, affective responses (*I love this!*) illustrated in Section 3.3.1 and the information-giving and storytelling presented in Section 4.3.1 suggest that some commenters did accept deepfake technology use for non-malicious purposes at face value. Many commenters discussed the technology as just another innovation to be critiqued and improved much the same as technology that came before, while others created videos to discuss intimate details of their family life.

Still, while Audience Theory and Encoding/Decoding can be used as a starting point to explore reactions to non-malicious deepfake technology use, conceptual gaps remain. First, given that deepfake technology is relatively new, there is a lack of empirical research to determine whether there is a correlation between (positively framed) exposure to deepfake technology and its general acceptance and use as a non-malicious technology (audience-as-outcome). Further, while this study showed that reception to non-malicious deepfake technology often changes in specific framings and contexts, the deepfake videos discussed were very specific to each case and may not speak to broader trends in the way people receive information through media.

5.4 Design Implications

This study examined creators' conceptualizations of their non-malicious deepfake technology use to develop videos and deepfake technology apps, as well as explored how people reacted to and interacted with these deepfake videos. Findings showed that despite creators intending for the technology and videos to be viewed as non-deceptive additions to the general tech and app landscape, in the absence of a clear information source, people sometimes struggled to understand what was real. Further, the use of peoples' images and photos without explicit consent, as well as the collection of public images removed from context, presents violations of contextual integrity. These findings suggest avenues for design.

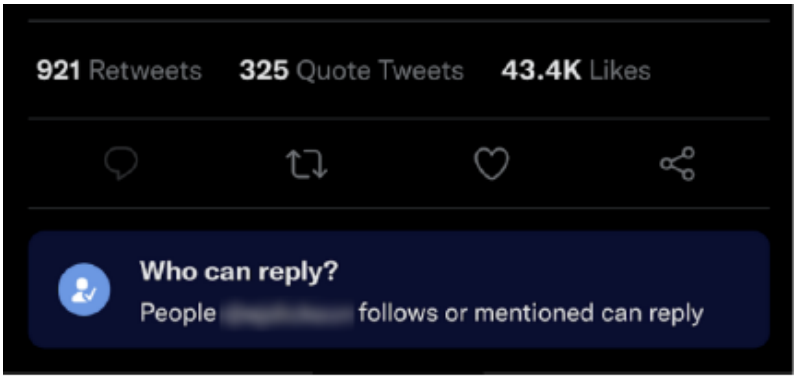
5.4.1 Affording Sensemaking

When reacting to the non-malicious deepfake technology (*deeptomcruise)* presented in Chapter

3, many commenters were confused, and in the absence of a clear signal, needed to rely on other commenters in the replies to make sense of what they saw. A design consideration that could address this would be for social media platforms to afford sensemaking; that is, to provide an easy means for people to determine or discover more about the video in an easily accessible way. Social media platforms already attempt to do this for harmful misinformation via *signifiers* in the form of labels, overlays, and accompanying text that provide additional context, and through specific colors that might signify a specific level of harm (Saltz et al., 2020; Sharevski et al., 2021).

Figure 7

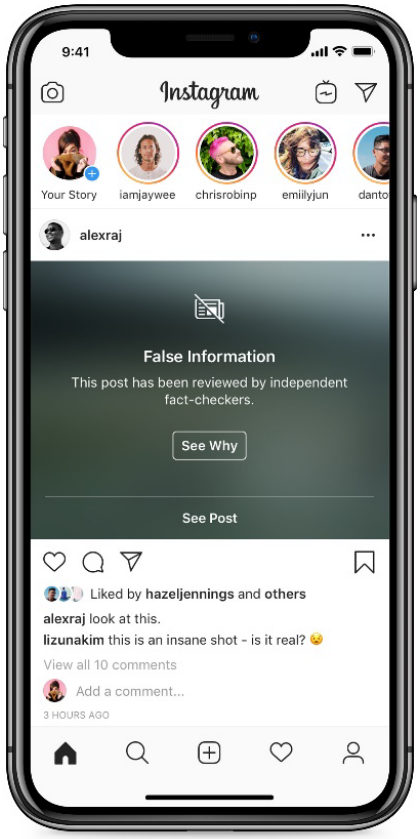
Twitter Label



Note. A label on Twitter indicates who can respond to a tweet. Screenshot taken by author.

Figure 8

Instagram False Information Label.



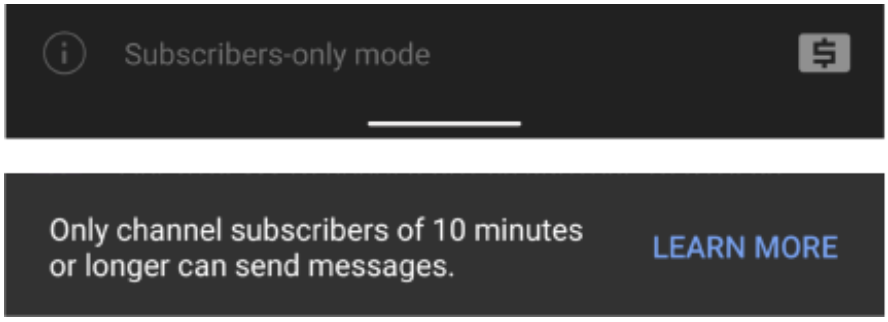
Note. Example of false information message taken from Instagram (Instagram Help Center, n.d.).

Recent research on the impact of warning labels on misleading Covid-19 information, however, has shown that, even though people believe it is the responsibility of platforms to ensure the information being shared is truthful and accurate, they often reject and even double-down on believing misleading information when presented with "paternalistic" warnings like labels and overlays (Saltz et al., 2020; Sharevski et al., 2021). However, non-malicious deepfake posts

might benefit from the use of an unobtrusive label or icon, color, or highlight that identifies their non-malicious nature. This would separate it from its more malicious and harmful counterparts (what Twitter terms misleading information/media (Gadde & Beykpour, n.d.), while still marking it as synthetic media. Unobtrusive signals are already well-represented on social and digital (video-sharing) platforms, affording specific functionality:

Figure 9

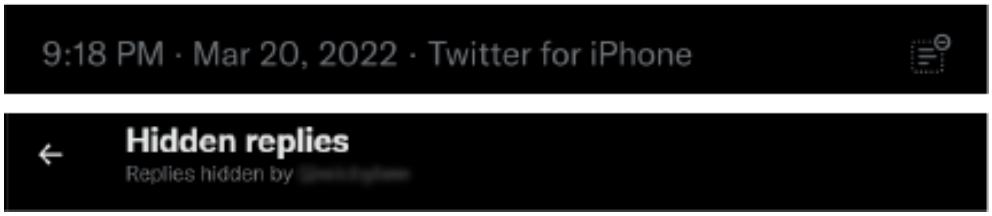
YouTube Information Labels.



Note: Clicking on the information icon on this YouTube alert shows “learn more” button. Screenshot taken by author.

Figure 10

Twitter Hidden Reply Icon



Note: Clicking on the hidden reply icon displays all replies that have been hidden by the tweet author. Screenshot taken by author.

User research could investigate peoples' use and reception of these signifiers and investigate if they view them in the same light as signifiers that identify malicious content.

5.4.2 Values and Ethics in Technology and Design

Unlike malicious deepfakes, non-malicious, creative deepfakes are not meant to be intentionally misleading. Further, the creator behind *deeptomcruise* and the app developers behind DeepNostalgia maintain they do not intend to promote or enact harm against those who view or generate deepfake videos using their technology. Still, my research discovered that a gap exists between what the creators and app developers intend (for non-malicious deepfakes to be considered in a similar light to other technical innovations that entertain and provide an emotional benefit) and peoples' impression that deepfake technology – even its non-malicious forms – are ethically ambiguous and may contribute to future harm.

Scholarship in legal studies and business has attempted to address what to do with malicious deepfakes, including promoting education and digital literacy, improving the technology that rapidly and accurately detects deepfakes, and creating laws and public policy prohibiting the development and use of malicious deepfake technology. United States Senator Sasse, for example, presented a bill to the U.S. Senate that proposes fines and possible imprisonment for those who create deepfakes that "facilitate criminal or tortious conduct under Federal, State, local, or Tribal law" (Malicious Deep Fake Prohibition Act of 2018, 2018). In another example, The R.E.A.L framework maintains that recording the original content (a "life alibi" that tracks all data about a person to ensure deniability of what the deepfake presents), exposing deepfakes early through technology via the use of detection software, advancing for legal protection against deepfakes, and leveraging trust between businesses and customers might address malicious deepfake concerns (Kietzmann et al., 2020). These frameworks and policy positions, however, address methods to prevent *malicious* deepfake technology. To address potential harm non-malicious deepfakes my present, an examination of ethical frameworks in design and technology may be more relevant.

Research in human-computer interaction has leveraged methodologies and practice from the social sciences, computer sciences, and information science to closely examine the ways technologists and designers think about ethical decision-making in their practice. Additional scholarship has examined the role that researcher intervention and embedding values advocates on technology teams as part of values-oriented design might promote ethical reflection, creativity, and the introduction of values levels (discussions about values in design) in practice

(Shilton, 2018). Through the process of ethical exploration, aligning design with ethics through ethical value quality requirements (EVRs), and promoting transparency and agreement about these requirements through the Ethical Value Register, Value-based Engineering (VBE) also encourages technologists and developers to think ethically about the products they develop, and might also promote systemic changes in work culture that also provokes deeper consideration and concern for how products may cause harm (Spiekermann & Winkler, 2020). Adopting a values-oriented/values-centered design or VBE approach to the development of deepfake technology apps might present a good first step to help creators and organizations address ethical concerns surrounding non-malicious deepfake technology development and use. A limitation to following a VCD or VBE approach is that these approaches focus on *teams* that have the means to include value experts as support and may not be as applicable in the case of hobbyists and individual creators.

5.5 Limitations and Future Research

5.5.1 Limitations

There were several limitations to this study. First, although I conducted a thorough examination of the posts, replies, conversations, and articles surrounding non-malicious deepfake technology use, explanations behind why commenters said and responded the way they did were rare. Reactions were also often emoji-heavy and needed to be excluded from the final data set due to my concerns with appropriate methodology and scope. And, although many responses were emoji coupled with text, emoji use, and meaning often differ from person to person. To address this, I relied on my own personal, subjective interpretations of what an emoji might mean

considering the accompanying text, which itself introduces bias. Finally, while a case-study approach was useful in understanding the uses of non-malicious deepfake technology and reactions to these videos, the findings may not be generalizable outside of the cases presented. Still, this study provides a solid foundation for understanding how prior research surrounding sensemaking, Audience Theory, and privacy as contextual inquiry might apply in the context of non-malicious deepfake technology reactions and use.

5.5.2 Future Work

To address issues of generalizability, future research could apply discourse analysis, sentiment analysis, content analysis, and other social-media analysis (SNA) methods as a complement to the qualitative research. Additional research might also include interviews with deepfake technology creators, technologists, and people reacting to deepfakes to explore their motivations and reactions to non-malicious deepfake technology. Additional qualitative and user research could also be used to examine people's reactions to signifiers (discussed in Section 5.4.1) that promote sensemaking. Additional research could also be conducted with technologists and creators who develop non-malicious deepfake technology to explore how they feel about the ethical implications of their work, and whether the adoption of an ethical framework might be beneficial in their practice.

Chapter 6: Conclusion

This study examined people's reactions to creative (non-malicious) deepfakes through analysis of two cases studies: the first examined replies and reactions to *deeptomcruise* on Instagram, while the second surveyed tweets and conversations surrounding the DeepNostalgia app. Using the theories and models of sensemaking, privacy as contextual integrity, and Audience Theory to guide my analysis, I found that in the absence of a clear information source, people used sensemaking and consulting strategies in attempts to understand the deepfake videos. I also found that although creators view their development of deepfake videos and apps as mostly benign, people are curious about the ethical implications inherent within the development and use of deepfake technology and see implications for future harm. I discuss avenues for social media platforms to use already implemented strategies (such as signifiers) to help people make sense of deepfake technology. Finally, I discuss how creators and developers might leverage methods from values-oriented/values-centered design and value-based engineering in design to guide their practice.

Appendices

Appendix A: Deeptomcruise Word Clouds

Word clouds generated from deeptomcruise posts 2-11 (see Table 1).

Word clouds generated from deeptomcruise posts 12-21 (see Table 1).

Appendix B: DeepNostalgia Word Clouds

Word clouds generated from each tweet batch (including initial 'pilot' data).

Appendix C: IRB Approval

1204 Marie Mount Hall
College Park, MD 20742-5125
TEL 301.405.4212
FAX 301.314.1475
irb@umd.edu
www.umresearch.umd.edu/IRB

DATE:	December 22, 2021
TO:	Keaunna Cleveland
FROM:	University of Maryland College Park (UMCP) IRB
PROJECT TITLE:	[1834403-1] An Exploration of Non-Malicious Deepfakes through Analysis of Two Case Studies
SUBMISSION TYPE:	New Project
ACTION:	APPROVED
APPROVAL DATE:	December 22, 2021
REVIEW TYPE:	Expedited Review
REVIEW CATEGORY:	Expedited review category # 5 & 7; Waiver of consent: 45CRF46.116(f)(3).

Thank you for your submission of New Project materials for this project. The University of Maryland College Park (UMCP) IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a project design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

Prior to final approval of this project scientific review was completed by the IRB Member reviewer.

This submission has received Expedited Review based on the applicable federal regulations.

This project has been determined to be a MINIMAL RISK project.

Please remember that informed consent is a process beginning with a description of the project and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the project via a dialogue between the researcher and research participant. Unless a consent waiver or alteration has been approved, Federal regulations require that each participant receives a copy of the consent document.

Please note that any revision to previously approved materials must be approved by this committee prior to initiation. Please use the appropriate Amendment forms for this procedure.

All UNANTICIPATED PROBLEMS involving risks to subjects or others (UPIRSOs) and SERIOUS and UNEXPECTED adverse events must be reported promptly to this office. Please use the appropriate reporting forms for this procedure. All FDA and sponsor reporting requirements should also be followed. All NON-COMPLIANCE issues or COMPLAINTS regarding this project must be reported promptly to this office.

Please note that all research records must be retained for a minimum of seven years after the completion of the project.

If you have any questions, please contact the IRB Office at 301-405-4212 or irb@umd.edu. Please include your project title and reference number in all correspondence with this committee.

Bibliography

Ahmed, S. (2021). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics*, *57*, 101508–101508. https://doi.org/10.1016/j.tele.2020.101508

Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, *97*(22), 3242-3250.

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research and Politics*, *6*(2). https://doi.org/10.1177/2053168019848554

Allyn, B. (2022, March 17). *Deepfake video of Zelenskyy could be 'tip of the iceberg' in Info War, experts warn*. NPR. Retrieved March 24, 2022, from https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

Amadoru, M., Fielt, E., & Kowalkiewicz, M. (2019). *Understanding Socio-cognitive Sensemaking of Digital Innovations in Twitter*.

Bode, L. (2021). Deepfaking Keanu: YouTube deepfakes, platform visual effects, and the complexity of reception. *Convergence: The International Journal of Research into New Media Technologies*, *27*(4), 919–934. https://doi.org/10.1177/13548565211030454

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

BuzzFeedVideo. (2018, April 17). *You won't believe what Obama says in this video!*. YouTube. Retrieved April 5, 2022, from https://www.youtube.com/watch?v=cQ54GDm1eL0&ab_channel=BuzzFeedVideo

de Ruiter, A. (2021). The Distinct Wrong of Deepfakes. *Philosophy & Technology*. https://doi.org/10.1007/s13347-021-00459-2

de Seta, G. (2021). Huanlian, or changing faces: Deepfakes on Chinese digital media platforms. *Convergence*, *27*(4), 935–953. https://doi.org/10.1177/13548565211030185

D'heer, E., & Verdegem, P. (2015). What social media data mean for audience studies: A multidimensional investigation of Twitter use during a current affairs TV programme. *Information Communication and Society*, *18*(2), 221–234. https://doi.org/10.1080/1369118X.2014.952318

Gadde, V., & Beykpour, K. (n.d.). *Additional steps we're taking ahead of the 2020 US election*. Twitter. Retrieved April 5, 2022, from https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes

Greengard, S. (2019). Will deepfakes do deep damage? *Communications of the ACM*, *63*(1), 17–19. https://doi.org/10.1145/3371409

Hall, S. (2000). Encoding/decoding. *Media studies: A reader*, 51-61.

Jia, Y. J., Chen, Q. A., Wang, S., Rahmati, A., Fernandes, E., Mao, Z. M., & Prakash, A. (2017, May). *ContexIoT: Towards Providing Contextual Integrity to Appified IoT Platforms*. https://doi.org/10.14722/ndss.2017.23051

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

Kerner, C., & Risse, M. (2021). Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. *Moral Philosophy and Politics*, *8*(1), 81–108. https://doi.org/10.1515/mopp-2020-0024

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

Klein, G., & Moon, B. (2006). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems*, *21*(4), 70–73. https://doi.org/10.1109/MIS.2006.75

Kunduru, S. R. (2018). *Social media and public discourse: A technology affordance perspective on use of social media features*. 168–176. https://doi.org/10.1145/3209626.3209627

Lee, Y. A., Huang, K. T. T., Blom, R., Schriner, R., & Ciccarelli, C. A. (2021). To Believe or Not to Believe: Framing Analysis of Content and Audience Response of Top 10 Deepfake Videos on YouTube. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 153–158. https://doi.org/10.1089/cyber.2020.0176

Limbong, A. (2020, December 16). *Difficulties of movie production during the pandemic*. NPR. Retrieved April 21, 2022, from https://www.npr.org/2020/12/16/947261145/difficulties-of-movie-production-during-the-pandemic

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

Marwick, A. E., & Boyd, D. (2014). Networked privacy: How teenagers negotiate context in social media. *New Media and Society*, *16*(7), 1051–1067. https://doi.org/10.1177/1461444814543995

Meskys, E., Liaudanskas, A., Kalpokiene, J., & Jurcys, P. (2020). Regulating deep fakes: Legal and ethical considerations. *Journal of Intellectual Property Law & Practice*, *15*(1), 24–31. https://doi.org/10.1093/jiplp/jpz167

Mihailova, M. (2021). To Dally with Dalí: Deepfake (Inter)faces in the Art Museum. *Convergence: The International Journal of Research into New Media Technologies*, *27*(4), 882–898. https://doi.org/10.1177/13548565211029401

Nissenbaum, H. (2004). *Privacy as Contextual Integrity* (Washington Law Review, Vol. 79, pp. 2–3). https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*, 119-157.

Öhman, C. (2020). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology*, *22*(2), 133–140. https://doi.org/10.1007/s10676-019-09522-1

Osborne-Gowey, J. (2014). What is Social Media? *Fisheries*, *39*(2), 55–55. https://doi.org/10.1080/03632415.2014.876883

Papacharissi, Z. (2016). Affective publics and structures of storytelling: Sentiment, events and mediality. *Information Communication and Society*, *19*(3), 307–324. https://doi.org/10.1080/1369118X.2015.1109697

Pérez Dasilva, J., Meso Ayerdi, K., & Mendiguren Galdospin, T. (2021). Deepfakes on Twitter: Which Actors Control Their Spread? *Media and Communication*, *9*(1), 301–312. https://doi.org/10.17645/mac.v9i1.3433

Pirolli, P., & Russell, D. M. (2011). Introduction to this special issue on sensemaking. *Human-Computer Interaction*, *26*(1–2), 1–8. https://doi.org/10.1080/07370024.2011.556557

Pu, J., Mangaokar, N., Kelly, L., Bhattacharya, P., Sundaram, K., Javed, M., Wang, B., & Viswanath, B. (2021). *Deepfake videos in the wild: Analysis and detection*. 981–992. https://doi.org/10.1145/3442381.3449978

Reducing the spread of false information on Instagram. Help Center. (n.d.). Retrieved March 24, 2022, from https://help.instagram.com/1735798276553028

Rini, R. (2020). *Deepfakes and the Epistemic Backstop*. www.philosophersimprint.org/020024/

Saltz, E., Leibowicz, C., & Wardle, C. (2020). *Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions*. http://arxiv.org/abs/2011.12758

Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2021). *Misinformation Warning Labels: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes*. http://arxiv.org/abs/2104.00779

Shilton, K. (2018). Values and ethics in human-computer interaction. *Foundations and Trends in Human-Computer Interaction*, *12*(2), 107–171. https://doi.org/10.1561/1100000073

Spiekermann, S., & Winkler, T. (2020). Value-based Engineering for Ethics by Design. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3598911

Stieglitz, S., Mirbabaie, M., Schwenner, L., Marx, J., Lehr, J., & Brünker, F. (2017). Sensemaking and Communication Roles in Social Media Crisis Communication. *Wirtschaftsinformatik und Angewandte Informatik*.

Thalen, M. (2022, March 16). *A deepfake of Ukrainian president Volodymyr Zelensky calling on his soldiers to lay down their weapons was reportedly uploaded to a hacked Ukrainian news website today, per @shayan86 pic.twitter.com/txlryecgy4*. Twitter. Retrieved April 5, 2022, from https://twitter.com/MikaelThalen/status/1504123674516885507

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media and Society*, *6*(1). https://doi.org/10.1177/2056305120903408

Vizoso, Á., Vaz-álvarez, M., & López-García, X. (2021). Fighting deepfakes: Media and internet giants' converging and diverging strategies against hi-tech misinformation. *Media and Communication*, *9*(1), 291–300. https://doi.org/10.17645/MAC.V9I1.3494

Webster, J. G. (1998). The Audience. *Journal of Broadcasting & Electronic Media*, *42*(2), 190–207. https://doi.org/10.1080/08838159809364443

Weick, K. E. (1995). *Sensemaking in organizations* (Vol. 3). Sage.

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11), 39–52. https://doi.org/10.22215/timreview/1282

Wu, L., Morstatter, F., Carley, K.M., & Liu, H. (2019). Misinformation in Social Media: Definition, Manipulation, and Detection. *SIGKDD Explor., 21*, 80-90.

Zhang, B., Zhou, J.P., Shumailov, I., & Papernot, N. (2020). Not My Deepfake: Towards Plausible Deniability for Machine-Generated Media. *ArXiv, abs/2008.09194*.

Zhang, P., & Soergel, D. (2014). Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking. *Journal of the Association for Information Science and Technology*, *65*(9), 1733–1756. https://doi.org/10.1002/asi.23125