Demystifying Deepfakes: An In-Depth Analysis of Technical Possibilities and Societal Implications

Robert Ville Schenk

Haaga-Helia University of Applied Sciences

Bachelor of Business Administration

Research-based thesis

2024

Abstract

Author
Robert Ville Schenk

Degree
Bachelor of Business Administration

Thesis Title
Demystifying Deepfakes: An In-Depth Analysis of Technical Possibilities and Societal Implications

Number of pages and appendix pages
28 + 8

This research-based thesis investigates the up-and-coming phenomenon of “deepfakes”, a term used to describe fake videos made using deep learning algorithms and face-swapping techniques.

Several academic papers and books serve as the basis for the research. Prominent cases of deepfakes are also described to explain the use cases and possibilities of deepfakes.

The research paper takes a qualitative approach, covering a wide range of topics, including ethics, legalities, and the algorithms behind deepfake software.

The thesis identifies the positive and negative possibilities of deepfakes and investigate technical aspects, such as specific algorithms used to make deepfakes and the technical requirements for generating credible deepfakes.

The motivations behind the people who create deepfakes differ significantly. As part of this study, five such content creators gave their input to answer questions about the risks, possibilities, and technical difficulty of generating deepfakes.

This research involves a practical part where an artifact was produced in the form of a deepfake featuring two prominent celebrities whose faces were swapped using the deepfake software DeepFaceLab. The artifact works as a proof-of-concept, and the process of creating the artifact can be followed in the thesis. This artifact was also evaluated by deepfake content creators on its quality and credibility.

Research questions are formed at the beginning of the study, and they guide the direction of the research. Conclusions to the research questions and the outcome of the study's findings are formulated in the end.

Keywords
Artificial Intelligence (AI), Deep Learning, Deepfake, General Adversarial Network (GAN), Neural Network

Table of contents

1 Introduction

In recent years, the landscape of digital content creation has undergone a transformative shift driven by advancements in deep generative approaches. One prominent outcome of this technological evolution is the emergence of Deepfakes, a term encapsulating the manipulation and generation of facial appearances with profound implications. This sophisticated technology has found applications ranging from harmless visual effect enhancements in the film industry to more nefarious activities, such as the propagation of misinformation through the fabrication of appearances of well-known individuals. (Juefei-Xu, Wang, Huang, Guo, Lei, Liu 2022, 1)

Deepfakes leverage Artificial Intelligence (AI) to generate entirely new video or audio content, aiming to portray events that never transpired. Artificial intelligence (AI) is the science of creating intelligent machines, particularly intelligent computer programs. It encompasses the development of systems that can perform tasks that typically require human intelligence, such as understanding language, critical thinking, and recognizing patterns. (IBM 2024)

Creating deepfakes involves various methods, with the most common relying on deep neural networks utilizing a face-swapping technique. Starting with a target video as the foundation, coupled with a collection of video clips featuring the person to be inserted, the AI program predicts and maps the individual's appearance onto the target. While the process is intricate, accessibility has increased with user-friendly applications like Zao and FakeApp, making basic deepfake generation feasible even for beginners. (Johnson & Johnson 2023)

1.1 Terminology

The term “deepfake” first appeared as the username of a Reddit user in 2017. The term became used to refer to pornographic videos using face-swapping technology but has since evolved to encompass various applications, including realistic still images and manipulated voice recordings. The term "artificial intelligence-generated synthetic media" is increasingly used to encompass a broader range of applications, including positive use cases in marketing and advertising. (Sloan 2020)

The "deep" in “deepfake” refers to the underlying technology, namely deep learning algorithms. These algorithms, through exposure to extensive datasets, solve complex problems, paving the way for the creation of synthetic content featuring real individuals. (Johnson & Johnson 2023)

The spectrum of deepfakes encompasses a range of manipulated media, varying in sophistication and the level of technical expertise required for their creation. (Paris & Donovan 2019, 10-14)

Here's an overview from less advanced to more advanced:

Traditional Editing

In the context of deepfakes, traditional edits are also referred to as cheap fakes. This category represent the less sophisticated end of the spectrum. They typically involve simple manipulations of existing content without leveraging advanced AI techniques. It includes basic video editing tools like cut-and-paste, speed alterations, or minor visual tweaks are employed to create deceptive content. (Paris & Donovan 2019, 10-14) For example, in 2020, the US Speaker of the House Nancy Pelosi appeared to be drunk in a video, when in reality the video was only slowed down. (Reuters 2020)

Generative Models

Definition: In this category, early generative models may be employed, but the technology is not as advanced as in full-fledged deepfakes. This means basic AI models, possibly using face-swapping algorithms, may be used for relatively simple manipulations. Some online tools or apps, for example Zao, allow users to perform basic face swaps or modify facial expressions. (Doffman 2019)

Deepfakes

Deepfakes represent a more advanced form of synthetic media using deep learning techniques. They involve deep neural networks, particularly Generative Adversarial Networks (GANs) to generate highly realistic content, such as facial expressions and lip movements. Its results are realistic face swaps, where one person's face is convincingly replaced by another in videos. (Paris & Donovan 2019, 10-14)

1.2 Impact

Since their inception, deepfakes have significantly impacted various industries, though primarily in malicious and non-consensual ways. In politics, they have been employed to defame and misrepresent prominent figures, such as the face-swapping of Argentina's former president, Mauricio Macri, with Adolf Hitler and the swapping of German Chancellor Angela Merkel's face with Donald Trump. In the realm of acting, deepfakes have contributed to movie productions, exemplified by the young face of Harrison Ford being deepfaked onto the character Han Solo in the 2018 film "Solo: A Star Wars Story." Furthermore, in the domain of fun and recreation, applications like Zao allow users to substitute their faces with those of celebrities or actors. However, the malicious potential of deepfakes is evident in their capacity to spread fake information and influence elections, even if

identified later, as the rapid circulation of such content can shape public opinion against the targeted individuals. (Khan, Bramah, Singh, Asthana 2023, 899)

Synthetic media, which deepfakes are part of, have already resulted in distressing consequences. In one case, a 14-year-old, Ellis, woke up to the circulation of AI-generated nude images of her on social media, causing significant emotional distress. The emotional toll on victims, including feelings of shame and fear, is evident, and the potential long-term consequences on individuals' mental health are underscored. The law has not kept pace with technology, leaving victims with limited recourse and raising concerns about the societal impact of deepfake proliferation. Cases like this emphasize the urgent need for legal frameworks to address the malicious use of deepfake technology, particularly in the context of non-consensual and harmful synthetic media. (STT–AFP 2023)

1.3 Regulation

The legal landscape around deepfakes is evolving due to their relatively recent emergence. The absence of clear legislation is evident, as platforms like Instagram hesitated to remove a deepfake video of Mark Zuckerberg in 2017, citing a lack of policies addressing deepfakes. Social media platforms have banned explicit content but haven't universally addressed non-consensual deepfakes. (Khan et al. 2023, 900)

While TikTok requires labels for realistic AI-generated content, some flagged videos lacked these disclosures. TikTok has taken action by removing or stopping the recommendation of accounts and videos violating policies related to posing as news organizations and spreading harmful misinformation. Additionally, a video using an AI-generated voice mimicking former President Obama was removed for violating TikTok's synthetic media policy. Despite TikTok's efforts, similar content has been found on platforms like YouTube, Instagram, and Facebook. The report underscores the challenge of dealing with AI-generated content, its potential to spread misinformation, and the limitations of user-selected labels. Various platforms have implemented measures, such as YouTube banning political ads using AI, Meta (Facebook) including an "altered" label in its fact-checking toolkit, and X (Twitter) requiring content to be significantly and deceptively altered to violate policies. (Thompson & Maheshwari 2023)

1.4 Research questions

How difficult is it to create a credible deepfake?

Is deepfake technology a danger to society?

What possibilities, positive and negative, does the development of deepfake technology open up?

How much of this topic is just hype?

These research questions collectively form a logical framework guiding the thesis forward by systematically addressing key aspects of the deepfake phenomenon. The first question delves into the technical intricacies, exploring the difficulty in crafting convincing deepfakes, thereby laying the groundwork for understanding the technology's complexity. The second question extends the inquiry to societal implications, probing the potential dangers posed by deepfake technology, emphasizing the broader consequences beyond its technical aspects. Moving forward, the third question broadens the scope to both positive and negative possibilities, exploring diverse applications and impacts of deepfake development. Finally, the fourth question encourages a critical evaluation of the discourse surrounding deepfakes to distinguish between sensationalism and genuine concerns. Together, these questions provide a systematic and holistic approach to understanding the complex nature of deepfake technology and its implications for society.

2 Research Method

The purpose of this qualitative study is to explore the landscape of deepfakes, focusing on their creation, technical aspects, and potential implications. The methodology involves a multifaceted approach, encompassing literature review, practical experimentation, and user input.

This study has a qualitative approach to gain in-depth insights into the phenomenon of deepfakes. Qualitative methods are well-suited for exploring complex and multifaceted topics, allowing for a nuanced understanding.

Extensive data was be collected through a thorough examination of books, research papers, and media sources. This literature review serves as the foundation for understanding the historical, social, and technical dimensions of deepfakes.

Research questions were formulated to guide the study. These questions address the motivations behind deepfake creation, the technical intricacies involved, and the potential societal impacts.

A proof-of-concept deepfake was created to gain firsthand experience with the technology. This practical step aims to further explain the complexities and challenges associated with deepfake generation. The technical details and steps involved in deepfake generation are explained, providing a comprehensive overview of the underlying technology.

Users of deepfake generation software were surveyed to gather their perspectives on the topic. Input was sought regarding general opinions on deepfakes, ethical considerations, and the skills required for creating deepfakes.

The resulting deepfake artifact was critically evaluated by deepfake generation software users based on quality and credibility. This evaluation serves to show the technical limitations of the technology.

Findings from the literature review, practical experimentation, and user input collectively contributed to answering the research questions. This holistic approach ensures a comprehensive exploration of the deepfake phenomenon.

3 Literature Review

This research is based on multiple books and academic papers. The literature review showcases multiple viewpoints of this topic from different authors. Case studies are also shown and analysed to show the uses of deepfakes so far.

3.1 Literary works

Deepfakes

Graham Meikle's book, titled "Deepfakes" and published in 2023, delves into the intriguing realm of synthetic media and its profound implications for our perception of reality in the digital age. As we grapple with the question of what happens when our visual trust is eroded, Meikle navigates the landscape of AI technologies capable of generating convincing videos depicting individuals saying and doing things they never uttered or performed. From the realms political satire to the intricacies of movie mashups and disinformation campaigns, the book deals with the concepts of trust and consent in the context of deepfakes. (Meikle 2023)

Meikle provides insights into the creation and utilization of deepfakes, shedding light on the evolving dynamics between remixing and sharing content and the diminishing capacity to trust the authenticity of media. With a focus on how deepfake videos offer a fresh perspective on the often-overlooked nature of contemporary media, the book explores the clash between our ability to remix and share content and our instinct to trust. Moreover, it examines the impact of these synthetic videos on the convergence of public and personal spheres in the social media environment, where all aspects of human experience become data for widespread sharing. (Meikle 2023)

Deepfakes aka Synthetic Media: Humanity at the Edge of an Uncanny Valley

This book, "Deepfakes aka Synthetic Media: Humanity at the Edge of an Uncanny Valley" by Ashish Jaiman, takes a comprehensive and multifaceted approach to the topic of deepfakes. The primary focus is on exploring the opportunities and challenges posed by synthetic media, particularly deepfakes, and the potential impact on individuals, society, democracy, journalism, and institutions. (Jaiman 2022)

The book is structured into five distinct parts, each contributing to a comprehensive exploration of the deepfake phenomenon. In the initial section, the author addresses the broader context of information disorder, discussing the prevalence of misinformation and proposing a cybersecurity-inspired approach to combat it effectively. The next section delves into the darker aspects of deepfakes, examining malicious use cases and the potential harms to individuals, society, democracy,

journalism, and institutions. The third part focuses on potential countermeasures, exploring individual, legislative, platform-based, and technical approaches to mitigate the risks associated with deepfakes. Shifting towards a more positive perspective in the fourth section, the book explores the constructive use cases of artificial intelligence and synthetic media, emphasizing their potential to empower individuals, contribute to equity, enhance education, and democratize entertainment. The final section delves into the ethical considerations surrounding deepfakes, providing technical insights into their design, implementation, and deployment, while emphasizing the crucial need for an ethical paradigm to guide the responsible development and use of synthetic media in the face of potential threats to democracy and society. (Jaiman 2022)

Deepfakes: A New Era of Misinformation

The research paper titled "Deepfakes: A New Era of Misinformation" authored by Rushan Khan, Bramah Hazela, Shikha Singh, and Pallavi Asthana explores the phenomenon of deepfakes from a technical and legal perspective. The paper shows technical aspects of deepfakes, offering formal definitions, an overview of their impact, the technical processes involved, and indicators to spot them. Additionally, the paper addresses the legal implications of deepfakes. The introduction underscores the need to educate the public about deepfakes, providing a detailed, technical, and easy-to-understand review. It also positions the paper as a resource for individuals interested in creating deepfakes, offering insights into the procedure, tools, and awareness on identifying deepfake videos. (Khan et al. 2023, 897-907)

3.1.1 Analysis of literature

The collection of literature for the thesis on deepfakes is comprehensive, covering diverse aspects of this emerging technology. Graham Meikle's book, "Deepfakes" (2023), provides an in-depth exploration of synthetic media, focusing on the erosion of visual trust and the profound implications for our perception of reality. Ashish Jaiman's book, "Deepfakes aka Synthetic Media: Humanity at the Edge of an Uncanny Valley" (2022), delves into the opportunities and challenges posed by synthetic media. The research paper titled "Deepfakes: A New Era of Misinformation" by Rushan Khan, Bramah Hazela, Shikha Singh, and Pallavi Asthana (2023) offers a focused examination of deepfakes from both technical and legal perspectives. The paper provides formal definitions, an overview of the impact of deepfakes, details on the technical processes involved, indicators for spotting them, and insights into the legal implications.

Together, these pieces of literature offer a comprehensive and nuanced view of the deepfake phenomenon, covering technical, ethical, legal, and societal dimensions. They collectively contribute to the depth and breadth of the thesis, providing a solid foundation for analysis and discussion.

3.2 Notable cases

The following cases highlight the use various use cases for deepfakes. They are prominent and impactful examples of how deepfakes have already been applied.

Case Mark Zuckerberg

A deepfake video featuring Facebook CEO Mark Zuckerberg surfaced on Instagram, created by artists Bill Posters and Daniel Howe along with the advertising firm Canny. The video (Figure 1), part of an exhibit called "Spectre" for the Sheffield Doc Fest, portrays Zuckerberg admitting to stealing user data. It underscores the potential misuse of deepfake technology. (Eadicicco 2019)

Figure 1. Deepfake of Mark Zuckerberg (Eadicicco 2019)

The deepfake video was made in response to Facebook's refusal to remove an altered video of Nancy Pelosi, artist Bill Posters sought to convey a message by creating a deepfake video featuring Mark Zuckerberg asserting the platform's control over its users. Intentionally provocative, the video aimed to highlight the ease with which politicians and public figures can be manipulated on social media through contemporary technology, encompassing various techniques beyond deepfakes. Despite its intent, the video's effectiveness was hindered by a mismatch in voices, underscoring the importance of skilled voice actors or advanced AI voice technology for the success of deceptive deepfakes. (Eadicicco 2019)

Case Kim Joo-Ha

In this case, a deepfake was used in a non-malicious manner for a unique and innovative application in television broadcasting. The MBN channel in South Korea employed a deepfake version of its regular news anchor, Kim Joo-Ha, to present the day's headlines. This deepfake, created by the South Korean company Moneybrain, aimed to replicate Kim Joo-Ha's voice, gestures, and facial

expressions with high fidelity (Figure 2). The broadcast was announced in advance to viewers, and the response was mixed. While some viewers were impressed by the realistic portrayal, there were concerns expressed about the potential impact on the real Kim Joo-Ha's job security. (Debusmann Jr. 2021)

Figure 2. Deepfake of Korean newsreader Kim Joo-Ha (Debusmann Jr. 2021)

MBN expressed its intention to continue using the deepfake technology for specific breaking news reports. Additionally, Moneybrain, the firm behind the artificial intelligence technology, revealed plans to explore opportunities with other media buyers in China and the United States. This instance demonstrates a non-malicious use of deepfake technology to enhance the presentation of news, showcasing its potential for creative applications in the media industry. (Debusmann Jr. 2021)

Case Harrison Ford

Solo: A Star Wars Story faced disappointment with both critics and box office performance, placing actor Alden Ehrenreich in the challenging position of filling Harrison Ford's iconic role as a younger Han Solo. YouTuber Shamook was not satisfied with Ehrenreich’s portrayal of Harrison Ford, so he made a deepfake of a younger Harrison Ford, achieving surprisingly convincing results (Figure 3). While typical deepfake videos often exhibit an uncanny valley effect, this one feels remarkably subtle and authentic, owing to Ehrenreich's existing resemblance to Han Solo. (Serrels 2020) The "uncanny valley" is a phenomenon where a human-like replica, such as a robot or computer-generated character, appears almost lifelike but still evokes feelings of unease or revulsion in observers. This occurs when the replica's appearance and behaviour are close to being realistic but still fall short of genuine human characteristics. (Paris 2022) Despite the absence of Han Solo's voice, the impact of seeing Ford's face in this context is notable. This sophisticated deepfake, although not

the first attempt, raises intriguing possibilities for experimenting with an alternate version of Solo: A Star Wars Story. (Serrels 2020)

Figure 3. Fan-made deepfake of Harrison Ford (Serrels 2020)

3.2.1 Analysis of cases

In the Case of Mark Zuckerberg, a deepfake video aimed to draw attention to the potential misuse of deepfake technology. The intentional provocation highlighted the ease with which public figures can be manipulated on social media. Despite its intent, the video's effectiveness was hindered, emphasizing the importance of skilled voice actors or advanced AI voice technology in the success of deceptive deepfakes. The Case of Kim Joo-Ha in South Korea illustrates a non-malicious application of deepfake technology in television broadcasting, showcasing the potential for creative applications in the media industry. The Case of Harrison Ford introduces a unique use of deepfakes in the film industry. Faced with the challenge of replicating Harrison Ford's iconic role as Han Solo, a YouTuber utilized deepfake technology to project Ford's face onto Alden Ehrenreich's body in "Solo: A Star Wars Story." This sophisticated deepfake achieved remarkably convincing results, opening intriguing possibilities for experimenting with alternate versions of movies.

The presented cases offer a comprehensive overview of different use cases of deepfakes, showcasing the versatility and potential impact of this technology across various domains.

4 Impact of Deepfakes

This chapter deals with the impact of deepfakes. The topic is approached from a legal and ethical perspective.

4.1 Survey

As part of this research, content creators on YouTube who specialize in deepfakes were approached. Nine individuals were contacted via email (appendix 1), resulting in five responses (appendix 2). The selection criteria prioritized individuals with technical expertise in deepfake technology. The people who were approached have YouTube channels where they showcase their work.

The following table is an overview of the survey participants.

Deepfake content creator	*Type of content creator*	*Country*	*Time since first deepfake posted*
No. 1	This content creator runs a company which sells deepfakes in the form of AI-generated personalized messages where celebrities are the models, and the user can decide which phrase the celebrity should say. The channel serves as promotion for the product.	United States	11 months
No. 2	This creator has posted hundreds of deepfakes in which faces of actors in films scenes where their faces are swapped with another actor.	United States	3 years
No. 3	This creator posts deepfakes of scenes from movies and tv shows where actors' faces are swapped or de-aged. The videos show a side-by-side view of the two different versions.	United States	10 months
No. 4	Similar to no. 2 and no. 3, this creator posts deepfakes where movies scenes are edited, but also deepfake versions of prominent public figures like Putin and Queen Elizabeth II.	United Kingdom	6 years

No. 5	This content creator mostly posts gaming videos but has posted a few deepfakes including a tutorial for the program DeepFaceLab.	Singapore	10 months

The purpose of reaching out to these content creators was to gather valuable insights from individuals actively engaged in the practical aspects of deepfake technology. Their firsthand experience was considered essential to enhance the depth and authenticity of the research. To acknowledge their time and contributions, respondents were offered a reward in the form of a USD 10 Amazon gift card or another payment method.

The survey questions were designed to elicit nuanced perspectives on various aspects of deepfake technology. The questions covered the perceived difficulty of creating credible deepfakes, opinions on the inherent dangers of deepfake technology to society, and reflections on the media portrayal of deepfakes, distinguishing between legitimate concerns and sensationalized hype. Additionally, respondents were invited to evaluate and provide feedback on a deepfake created as part of the study. The survey aimed to gather diverse opinions and insights from individuals actively involved in the deepfake creation community.

4.2 Use cases of deepfakes

Deepfakes are produced by a diverse range of actors, categorizable into four main types. First, there are deepfake hobbyist communities, comprised of individuals who engage in creating deepfakes for entertainment, often focusing on meme-like or humorous content. These hobbyists, initially drawn to the technology as a form of online humour, have formed communities where they share their creations. The motivation for these hobbyists is typically centred around solving intellectual puzzles and contributing to the development of AI-generated videos. (Westerlund 2019, 41-42) Four out of five survey respondents in this study are part of the first category of deepfake program users.

Second, political players, including foreign governments, activists, and various political agitators, constitute another category of deepfake producers. These actors leverage deepfakes as tools in disinformation campaigns aimed at manipulating public opinion, undermining confidence in institutions, and interfering with elections. In the context of hybrid warfare, deepfakes become weaponized disinformation tailored to exploit specific social media users' biases. (Westerlund 2019, 42)

Third, malevolent actors, such as fraudsters, represent a group that employs deepfakes for malicious purposes, including market and stock manipulation, financial crimes, and impersonation. Criminals have used AI-generated fake audios for phone-based impersonation, attempting to deceive individuals into making urgent cash transfers. (Westerlund 2019, 42-43)

Finally, legitimate actors, such as television companies, also engage in deepfake production, often collaborating with hobbyist communities. While hobbyists focus on entertaining content, television companies may utilize deepfake technology for more professional purposes, such as creating music videos or enhancing television shows. (Westerlund 2019, 41-43) One respondent of the survey in this study represents a company which creates and distributes deepfake edits. They can be considered part of this group because it is a company creating deepfakes for commercial purposes.

Despite the negative associations commonly attached to the term "deepfakes," which typically elicits concerns about deception, the technology is experiencing a growing presence in commercial applications. Referred to more formally as AI-generated videos or synthetic media, this technology is making significant strides in sectors like news, entertainment, and education. Synthesia, a London-based company, exemplifies this trend by creating AI-powered corporate training videos for global entities such as WPP and Accenture. The CEO of Synthesia, Victor Riparbelli, highlights the ease with which global firms can leverage AI-generated videos to communicate complex information in multiple languages, streamlining processes that would otherwise involve extensive recording efforts. This underscores the positive impact of deepfake technology in facilitating efficient and cost-effective content creation for legitimate commercial purposes. (Debusmann Jr. 2021)

Another company that offers deepfakes is LipSynthesis. It introduces an innovative application that utilizes advanced deepfake technology and natural language processing (NLP) to create realistic videos of individuals delivering specified text. The technology finds application in both entertainment and educational contexts. In practise, this means that you can choose a celebrity and decide what they should say in the video. The platform emphasizes responsible usage, with guidelines prohibiting commercial purposes, defamatory language, and requiring clear indicators for synthetic media. (LipSynthesis.com 2024)

4.3 Privacy and intellectual property

The website thispersondoesnotexist.com uses artificial intelligence (AI) to generate realistic-looking faces of people who do not actually exist. The website continually generates new faces each time you visit, showcasing the capabilities of AI in creating lifelike and convincing images (Figure 4). It is an example of how advanced AI algorithms can simulate the appearance of human faces with remarkable accuracy. (Vincent 2019)

Figure 4. AI-generated faces (thispersondoesnotexist.com)

The website uses images from the image-sharing website Flickr as source material for its algorithm. The images were used to train the algorithm without explicit consent from users. Facial images can be considered sensitive personal information and using them without proper consent may violate privacy norms. Even if the images were publicly available on platforms like Flickr, individuals may not have anticipated that their photos could be used in this manner. Most of the original photos were uploaded in the 2000s when Flickr was at the peak of its popularity, meaning that users could not have anticipated or consented to their photos being used in this way. (Meikle 2023, 34-35)

The creation of deepfakes involves the use of personal data, subject to the regulations laid out in the General Data Protection Regulation (GDPR). Personal data encompassing voice fragments, photos, and videos are utilized throughout the deepfake lifecycle. GDPR's broad definition of "processing" applies to technology developers and creators, requiring compliance in the creation and dissemination of deepfakes. Legal grounds for processing personal data include 'informed consent' and 'legitimate interests,' with the latter potentially applicable in cases such as satirical deepfakes. When legitimate interests are not relevant, obtaining informed consent from individuals featured in both the original and fabricated videos is mandatory to comply with GDPR. Non-compliance risks GDPR violations. These requirements do not extend to deepfakes of deceased individuals, but specific laws at the Member State level may necessitate obtaining consent from heirs. GDPR offers mechanisms for victims to address unlawful deepfake content, including the right to correct or delete inaccurate data. However, legal recourse for victims can be challenging due to anonymity and resource constraints. (European Parliament Research Service (EPRS) 2021, 38-39)

The source material for deepfakes typically includes images and videos that may be copyrighted or owned by someone else. As a result, the creation, distribution, and use of deepfakes can potentially infringe on existing intellectual property rights. Additionally, establishing responsibility for malicious use or harm caused by deepfakes further complicates the legal landscape. The evolving technology and lack of clear legal precedents make it difficult to address intellectual property

issues related to deepfakes, creating a complex and challenging environment for regulation and protection. (Meikle 2023, 34-35)

4.4 EU laws and regulations

The European Union (EU) has taken an active approach to regulating deepfakes, highlighting the importance of thorough research into detection and prevention methods, while also suggesting rules for clearly labelling artificially generated content. Various regulatory frameworks within the EU, including the AI regulatory framework, General Data Protection Regulation, Copyright regime, e-Commerce Directive, Digital Services Act, Audio Visual Media Directive, Code of Practice on Disinformation, Action plan on disinformation, and Democracy action plan, cover different aspects of deepfake governance. Notably, the EU has introduced legislative measures mandating social media platforms to eradicate deepfakes and disinformation. The Code of Practice on Disinformation, updated in June 2022, now includes fines of up to 6 percent of global revenue for failure to comply, strengthening its impact. Initially voluntary, this code gained more authority with the backing of the Digital Services Act, which became effective in November 2022, enhancing oversight of digital platforms. Furthermore, the proposed EU AI Act outlines transparency and disclosure requirements for providers of deepfakes. (Lawson 2023)

The European Commission proposed a comprehensive “AI regulatory framework” in April 2021, aiming to ensure trustworthy and secure applications of AI while respecting the values and fundamental rights of EU citizens. The framework adopts a risk-based approach, categorizing AI systems into 'unacceptable risk,' 'high risk,' 'limited risk,' and 'minimal risk.' For deepfake technologies, the proposal allows usage but imposes transparency obligations, particularly in labelling content to indicate artificial manipulation. However, the nature and scope of this measure raise uncertainties, lacking concrete guidelines for disclosure and penalties for non-compliance. The proposal acknowledges exceptions for authorized uses, such as law enforcement purposes or the exercise of freedom of expression and the arts. The effectiveness of these requirements in deterring malicious actors remains uncertain, prompting the need for further enhancements in the regulatory framework. (European Parliament 2023)

4.5 Regulation in other countries

United States

In the United States, deepfake legislation began at the state level, with California and Texas leading the way in 2019. California's AB 730 prohibits the distribution of manipulated content featuring political candidates within a 60-day period before an election, similar to Texas's law but with a 30-day restriction. Criticism has been voiced, especially regarding potential infringements on free

speech. Federal measures, such as the Identifying Outputs of Generative Adversarial Networks Act (IOGAN Act) and the U.S. National Defence Authorization Act, focus on information collection, research, and standards to combat deepfakes. (European Parliament Research Service (EPRS) 2021, 45)

India

India lacks specific deepfake legislation but is considering amendments to data protection and copyright laws. Discussions highlight the need for social media platforms to address deepfakes more rigorously. Unique to India is the focus on posthumous deepfakes, with suggestions to supplement data protection laws for the protection of deceased individuals' images. (EPRS 2021, 46)

China

Deepfakes are prevalent in China, prompting the government to enact legislation effective from January 1, 2020. The law mandates labelling of deepfake content by app providers, with strict rules against the production and dissemination of fake news. Chinese authorities use multiple measures, including user registration with identifiable information, easy complaint channels and regular inspections to enforce the regulations. (EPRS 2021, 46)

Taiwan

While Taiwan lacks specific deepfake legislation, its strategy against fake news, influenced by China's disinformation efforts, involves a 'nerd immunity' approach. This strategy employs professional fact-checkers and trains the public to recognize and counter false information actively. Citizens play a pivotal role in actively combating disinformation by disseminating corrective content creatively. This approach offers valuable insights into the EU's ongoing debate on addressing deepfakes. (EPRS 2021, 47)

4.6 Societal impact of deepfakes

Deepfakes represent a significant facet of a broader trend on the internet, reflecting the intricate relationship between user-generated content, privacy concerns, and the evolving landscape of digital manipulation. As users share personal data and content with large corporations, there is an implicit understanding that this information might be utilized for various purposes beyond its initial context. This phenomenon reflects a shift in the perception of private life, where individual data points become integrated into the public domain. (Meikle 2023, 7-8)

Similarly, the proliferation of deepfakes aligns with this paradigm, as individuals willingly contribute to the vast pool of online content, including photos and videos. The emergence of deepfake

technology allows for the manipulation of this shared visual material, blurring the lines between authenticity and manipulation. The implicit consent to the potential use and manipulation of personal content becomes an inherent part of the digital experience. In the realm of deepfakes, the act of posting photos and videos online exposes individuals to the risk of having their visual identity altered without consent. (Meikle 2023, 7-8)

When consuming media in the news or online, people now have to consider that even videos can be fakes using deepfakes. This makes people more sceptical of any content they see, no matter if real or not. It also makes it easier for anyone to claim that a video is a deepfake, just because it is more convenient to them. When a video surfaced in 2016 of Donald Trump where he brags about how he mistreats women, he first defended himself saying that it was “locker room talk”. Later, he claimed on numerous occasions that the video was faked. The existence of deepfakes and the healthy scepticism people have towards the authenticity of media content is enough for someone to take advantage of this uncertainty and claim an inconvenient truth to be a lie. This is called the liar’s dividend. (Jaiman 2022, 16)

5 Technical Side of Deepfakes

Creating high-level deepfakes is an intricate process. Behind deepfake software lie complex algorithms. The technical details of generating deepfakes are explained in this chapter. A deepfake featuring two celebrities is created and the process is shown here.

5.1 Algorithms involved in deepfake generation

Neural Network

A neural network is a computational model which mimics the structure and processes of the human brain. It is composed of interconnected nodes, or artificial neurons, organized into layers. Just like a brain has neurons, these nodes work together to process information. (Yasar 2023)

The layers of a neural network can be imagined as team of workers. The first team (input layer) receives a task, passes it to the middle team (hidden layers) that does some thinking, and then the last team (output layer) gives the final answer.

Each worker (node) in the team has its own job, and they talk to each other through connections. These connections have strengths, like how much one worker trusts another. The team gets better over time by adjusting these trust levels based on whether they did a good or bad job.

The goal is to train the team so that when they get a new task (input), they give the correct answer (output). This helps in solving various problems, like recognizing pictures, understanding speech, or making predictions.

So, a neural network is basically a smart system that learns by doing tasks over and over, adjusting itself to get better at those tasks. (IBM 2023)

Figure 5 shows a visual representation of the various layers of a neural network.

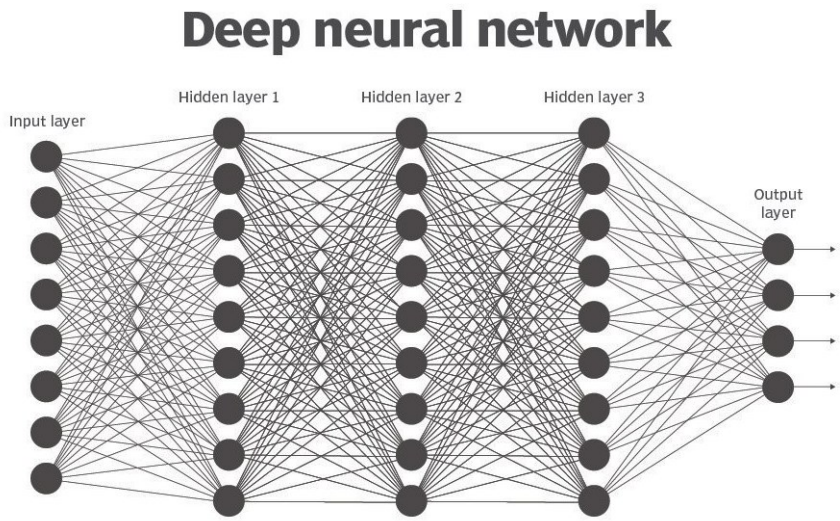

Figure 5. (Deep) neural network visualisation (Yasar 2023)

Neural networks are the underlying part of deep learning, a subset of machine learning. Deep neural networks, or deep learning models, consist of multiple hidden layers, enabling them to learn intricate patterns and representations from large datasets. Neural networks are applied in various domains, including image and speech recognition, natural language processing, and many other tasks where pattern recognition and complex data relationships are crucial. (Yasar 2023)

Neural networks play a central role in the creation of deepfakes. Deepfakes use complex artificial intelligence algorithms, often based on deep neural networks, to manipulate and generate highly realistic fake videos or images. These networks analyse and learn patterns from large datasets, enabling them to mimic the appearance and behaviour of a person in existing media. By training on diverse facial expressions, movements, and voices, neural networks can create convincing simulations, making deepfakes a product of advanced machine learning technologies. (Meikle 2023, 32)

General Adversarial Network

Generative Adversarial Networks (GANs) represent a category of artificial intelligence algorithms utilized within unsupervised machine learning frameworks. They were introduced by Ian Goodfellow and his colleagues in 2014. GANs are designed to generate new, realistic data instances that resemble a given dataset. The key innovation of GANs lies in their generative nature, as they create new content rather than simply classifying or recognizing existing patterns. (Meikle 2023, 34-35)

The GAN structure consists of two neural networks, called a generator and a discriminator, which are trained in parallel through so-called adversarial training. (Hansen 2022)

The idea behind GANs is like a creative game between two teams: a generative team and a discriminative team.

The generative team's job is to produce something new, like creating fake images. The key is to make these fake images look as realistic as possible.

On the other side, the discriminative team acts like the judge. Their task is to figure out which images are real and which ones are fake. They use a model called a discriminator for this. The game becomes a competition, pushing both teams to get better at their jobs.

The interesting part is that as the generative team gets better at creating realistic fakes, the discriminative team also improves at telling the difference. This back-and-forth makes the generated images look more and more like the real thing. Goodfellow's analogy likens it to a team of counterfeiters making fake money and the police trying to catch them. The competition between them makes both sides better until it's hard to distinguish between the real and the fake. (Hansen 2022)

Figure 6 shows a visual representation of the tasks which the discriminator and generator perform in a GAN.

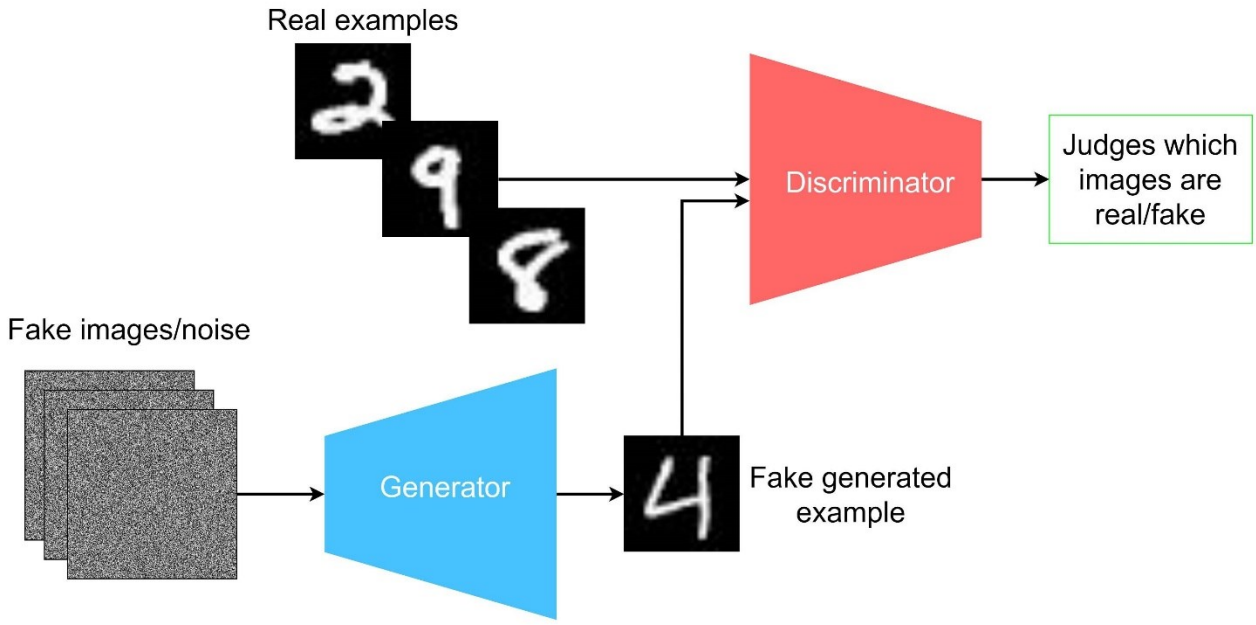

Figure 6. Visualisation of GAN (Hansen 2022)

It's a powerful tool as long as you have good data to train the model. (Hansen 2022)

GANs are used when making deepfakes. The combination of generative and discriminative models results in the generation of highly convincing fake media. The generative model keeps improving

its ability to create realistic content, while the discriminative model becomes adept at distinguishing between real and fake. This constant competition leads to the creation of deepfake videos or images that are challenging to differentiate from authentic ones. (Meikle 2023, 33-35)

5.2 DeepFaceLab

DeepFaceLab is a prominent program in the realm of deepfake technology, focusing on the generation of deepfakes. It is a free and open-source software with a text-based interface (Figure 7). This framework equips users with essential tools for the creation of high-quality face-swapping content. It makes it possible to achieve cinema-quality results with a high level of fidelity. According to the community forum DeepfakeVFX.com, more than 95% of deepfake videos are created with DeepFaceLab. (DeepfakeVFX.com 2021)

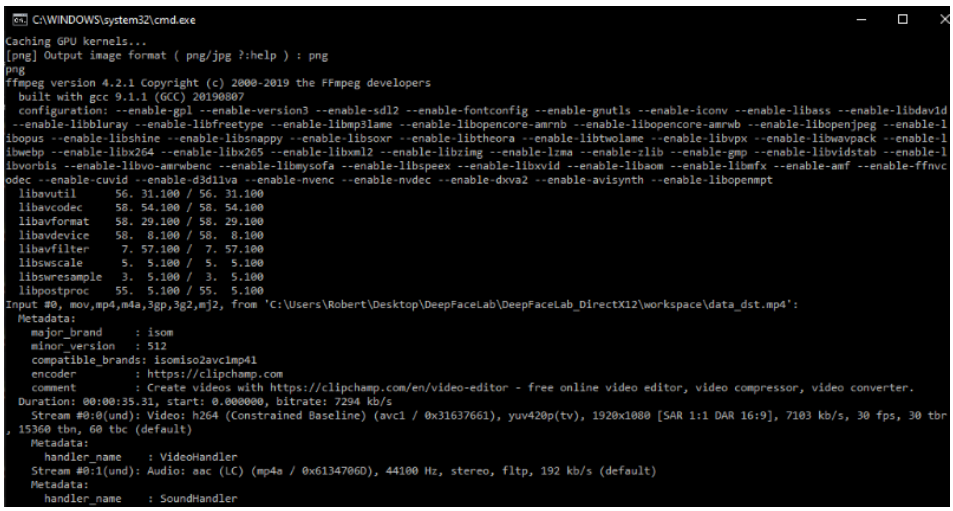

Figure 7. Screenshot of the interface of DeepFaceLab

5.3 Process of creating deepfake

For this research paper, a deepfake was created using DeepFaceLab. The original video onto which was used is a clip of Dua Lipa giving an interview. The aim was to project Anne Hathaway's face onto this video as a deepfake. Using DeepFaceLab, the frames from the video were extracted, identifying the face in each frame. Figure 8 shows a frame of the video along with the outlines of the face which the program identified.

Figure 8. Frame from original video of Dua Lipa interview (left) and facial recognition markings (right)

Next, a face set of Anne Hathaway was downloaded from DeepfakeVFX.com, featuring images contributed by a forum member. This face set comprised around 10 000 photos of Anne Hathaway, sourced from various videos. Additionally, a pre-trained model form DeepfakeVFX.com was used which is a compilation of numerous faces displaying different angles and expressions, serving to educate the AI on general facial characteristics.

The central aspect of the process involved merging Anne Hathaway's face with the facial expressions of Dua Lipa. The interface, as illustrated in Figure 9 and Figure 10, displayed images of Anne Hathaway and Dua Lipa alongside the AI's interpretations. The final image on the right depicts how the program envisions Anne Hathaway's face with Dua Lipa's expressions. The merging process was time-intensive, requiring approximately 25 000 iterations for Anne Hathaway's face to become recognizable. Due to a moderately powered graphics card, the duration of iterations averaged one second, ranging between 0.5 and 3 seconds. The accompanying yellow graph in Figure 9 and Figure 10 illustrates the iteration duration, showing increased efficiency over time as the AI familiarizes itself with the facial features.

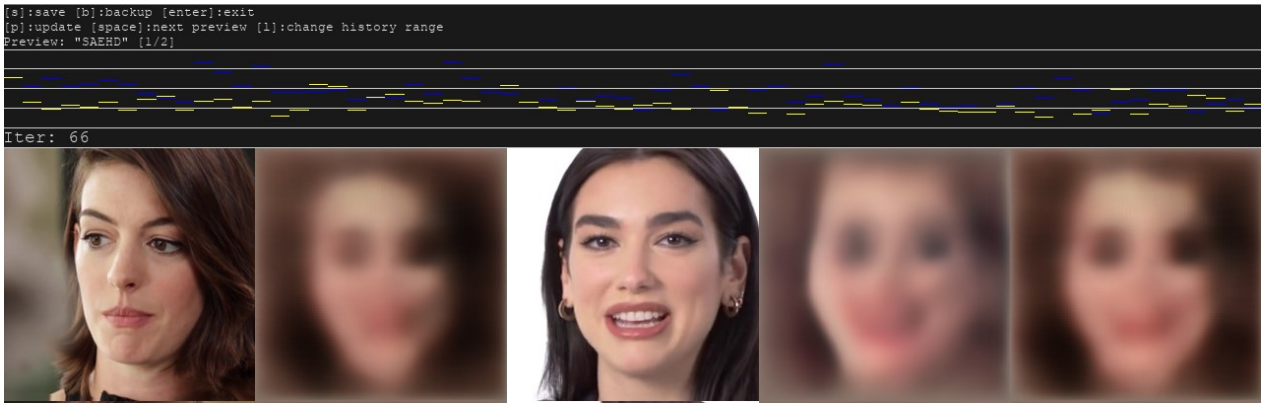

Figure 9. Faces generated after 66 iterations

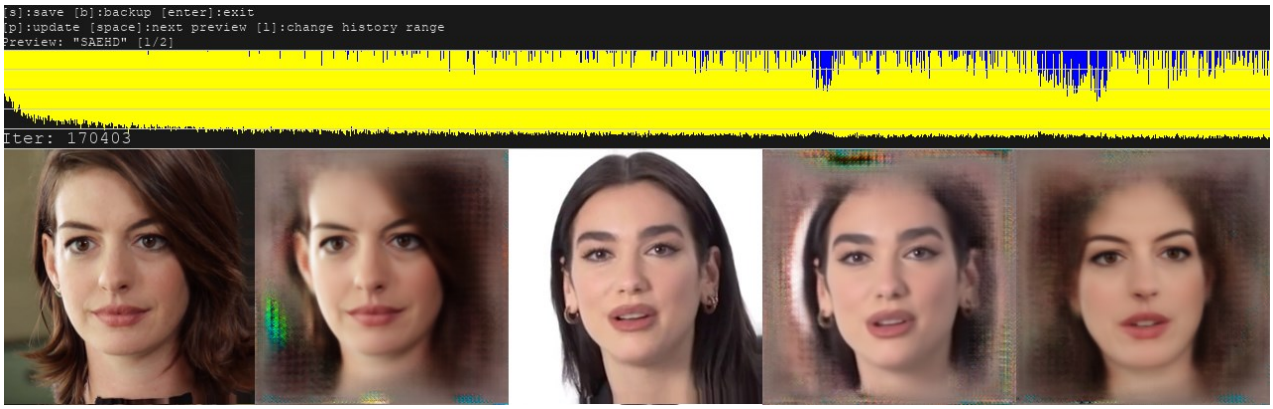

Figure 10. Faces generated after 170 403 iterations

The process spanned 55 hours, running on various settings throughout a week and totalling around 175 000 iterations. It is up to the user to decide how many iterations are necessary. The results of this deepfake creation are presented below (Figure 11 and Figure 12).

Lastly, the generated face was merged with the original video. There are many blending options, and it took two attempts and feedback from other DeepFaceLab users to get a good result.

The following (Figure 11) is a frame from the first attempt at rendering the deepfake.

Figure 11. Screenshot of first attempt at generating deepfake

The video is available here:

https://youtu.be/GtFSXP9VpPI

The clip was sent to content creators on YouTube who post deepfakes. They were approached by email (Appendix 1) and offered a $10 Amazon gift card to give feedback on the deepfake in Figure 11. The responses (Appendix 2) contained the following feedback and suggestions, among others:

Evaluation

- Good first effort
- Low quality result
- Deepfake not convincing

Suggestions

- Use a higher resolution pre-trained model
- Prepare face set more
- Increase size of face mask and blur edges
- For blending, use "overlay" mode instead of "seamless" mode
- Shorten the video to reduce processing material
- Significantly longer processing duration
- Implement proper post-editing to address artifacts/glitches, especially when hands cover the face

After analysing the responses, the deepfake was processed further and rendered using the different blending setting in the final stage. The following image (Figure 12) shows a frame from the final deepfake result:

Figure 12. Screenshot of second attempt at making deepfake

More screenshots from the deepfake are in appendix 3. The full video is available here:

https://youtu.be/TPQMvAYCf9c

5.4 Difficulty

The interface of DeepFaceLab interface is not very user-friendly, as it is text-based. There is no official guide on how to use it, only a few tutorials on YouTube uploaded by users of the program. Additionally, the community of users is rather small. There is a forum with around 15 000 active members. When encountering technical problems, the software gives vague error messages.

While making the deepfake, an issue arose when setting up the program to merge the faces. The error message “Check failed: ptr != nullptr Invalid pointer” was shown.

The error message was not of much help there was only one related post on the community forum DeepfakeVFX.com where a user suggested that it might have something to do with the GPU’s (Graphics Processing Unit) insufficient processing power. There was no suggestion for how to fix it. After a lot of hassle, the program managed to work by putting the processing settings on a lower intensity and processing some tasks on the CPU (Central Processing Unit) instead of the GPU (Graphics Processing Unit).

According to the deepfake software users who participated in this research, there is a steep learning curve to deepfakes. It takes time and a powerful computer. There are many things, such as the use of pretrained models to make the process faster. Since the technology is available for free, it is technically possible for anyone to make deepfakes.

6 Conclusion

6.1 Difficulty of generating a credible deepfake

The process of generating a credible deepfake involved several factors that contributed to its success. Firstly, the abundance of images of Anne Hathaway helped the training of the model, greatly improving the authenticity of the final result. Additionally, the use of a moderate PC, coupled with patience and determination, played a crucial role in overcoming the technical challenges associated with deepfake creation. Support and guidance from experienced deepfake content creators further helped the process, providing valuable insights and assistance.

Moreover, the resemblance between Anne Hathaway and Dua Lipa, particularly in hairstyle, added to the convincing nature of the deepfake. It's important to note that the deepfake generated as part of this study did not involve modifying the audio or the context of the content, serving primarily as a proof of concept rather than something to be used for nefarious or otherwise impactful uses.

Insights from survey responses provided by deepfake content creators shed light on the perceived difficulty of creating credible deepfakes. While the process itself may not be inherently difficult for someone with moderate computer literacy, it is time-consuming and resource intensive. The availability of open-source programs like DeepFaceLab simplifies the technical aspects, but the significant investment of time, computational power, and data is required to achieve credible results. Deepfake content creators emphasized the importance of data quality, hardware specifications, and the learning curve associated with deepfake creation.

When creating the deepfake which is part of this study, the text-based interface of DeepFaceLab posed usability challenges, compounded by the lack of comprehensive documentation and limited user community support. Technical errors encountered during the process highlighted the complexities of configuring the software and troubleshooting hardware compatibility issues.

While advancements in deepfake technology have made it increasingly accessible to everyday users, the process of generating credible deepfakes remains arduous and technically demanding. It requires a combination of technical expertise, computational resources, and perseverance to navigate the complexities of deepfake creation effectively.

6.2 Risk to society

Deepfakes pose several theoretical scenarios and potential dangers, raising concerns in various domains such as cybersecurity, misinformation, privacy, and even national security. One significant risk is the potential for deepfakes to be used in spreading disinformation and fake news. Malicious

actors could create realistic videos of public figures saying or doing things they never did, leading to confusion, damage to reputations, and even political manipulation. (Jaiman 2022, 88-89)

The proliferation of deepfakes poses significant risks to society. Deepfakes have been utilized in political spheres to defame and misrepresent prominent figures, exemplified by the face-swapping of political leaders such as Mauricio Macri and Angela Merkel with controversial historical figures. Such manipulations not only erode trust in political discourse but also have the potential to influence public opinion and electoral outcomes (Khan et al. 2023, 899).

In addition to political manipulation, deepfakes have been instrumentalized for non-consensual and malicious purposes, leading to distressing consequences for individuals. Cases like the circulation of AI-generated nude images of a 14-year-old victimize individuals, causing significant emotional distress and long-term psychological consequences (STT–AFP 2023). The emotional toll on victims underscores the urgent need for legal frameworks to address the malicious use of deepfake technology, particularly concerning non-consensual and harmful synthetic media.

Furthermore, deepfakes have been weaponized as tools in disinformation campaigns, perpetrated by political players, foreign governments, activists, and malicious actors. These campaigns aim to manipulate public opinion, undermine confidence in institutions, and interfere with elections (Westerlund 2019, 41-44). The rapid circulation of such disinformation, facilitated by the internet and social media platforms, poses a significant threat to democratic processes and societal stability.

Moreover, the misuse of personal data in deepfake creation raises profound privacy and intellectual property concerns. The use of images and videos without explicit consent violates privacy norms and infringes on intellectual property rights, highlighting the need for robust legal frameworks and regulatory measures to address these challenges (EPRS 2021, 44-45).

6.3 Possible positive outcomes of deepfakes

Deepfakes have enabled remarkable feats in the entertainment industry, offering solutions to challenges such as character aging or de-aging and resurrecting performances by deceased actors. For instance, the incorporation of Harrison Ford's iconic face onto Alden Ehrenreich's body in "Solo: A Star Wars Story" showcased the potential of deepfake technology to seamlessly blend visual elements in film production (Serrels 2020).

Companies like Synthesia and LipSynthesis are leveraging deepfakes for legitimate commercial purposes, such as creating multilingual corporate training videos and facilitating personalized video content (Debusmann Jr. 2021; LipSynthesis.com 2024). Such examples highlight the potential of deepfakes to revolutionize content creation and communication in various industries.

The survey (appendix 2) showed that while some argue that deepfake technology holds potential for beneficial applications, such as enhancing movie production or enabling digital avatars for anonymity, the darker side of deepfakes cannot be overlooked. The ease with which convincingly realistic videos can be produced exacerbates the challenge of distinguishing between authentic content and sophisticated fabrications, undermining public trust in media and information sources.

6.4 Media attention and hype

The media's coverage of deepfakes sparks a range of responses from the deepfake content creators involved in this study, with opinions varying from genuine concern to scepticism about sensationalism. Some highlight the potential for deepfakes to deceive and manipulate, particularly among older generations less familiar with the technology's capabilities. Others caution against conflating different forms of AI and emphasize the benign uses of deepfake technology. However, there's a consensus among deepfake content creators that media coverage often lacks depth and understanding of the intricacies involved. Despite this, there's acknowledgment of the importance of raising awareness about deepfakes, even if it comes at the cost of occasional sensationalism. Ultimately, there's a call for more nuanced and balanced reporting to inform the public effectively.

6.5 Future research opportunities

The topic of deepfakes can be studied in more detail in a few ways. It would be interesting to investigate the credibility of deepfakes in a study where participants are asked to point out which video is fake. It can also be studied if seeing a deepfake of a public figure saying something shocking influences the viewers' perception of them as a person, even if they are informed that it is a deepfake.

Another topic which has not been assessed in this study is prevention and possible solutions to problems that are brought about by deepfakes. This means investigating deepfake detection software and possible legal solutions to the ethics and privacy issues mentioned in this study.

Deepfakes are a relatively new phenomenon and the public discourse around the topic will likely evolve, possibly opening up new topics and nuances which are worthy or research.

Sources

Debusmann B. Jr. 2021. 'Deepfake is the future of content creation'. BBC News. Available at: https://www.bbc.com/news/business-56278411 (accessed 16 Dec 2023).

DeepfakeVFX.com. 2021. DeepFaceLab. Available at: https://www.deepfakevfx.com/downloads/deepfacelab/ (accessed 4 Jan 2024).

Doffman, Z. 2019. Chinese Deepfake App ZAO Goes Viral, Privacy Of Millions 'At Risk'. Forbes. Available at: https://www.forbes.com/sites/zakdoffman/2019/09/02/chinese-best-ever-deepfake-app-zao-sparks-huge-faceapp-like-privacy-storm/?sh=1fd5f0f38470 (accessed 6 Feb 2024).

Eadicicco, L. 2023. Deepfake video on Mark Zuckerberg surfaces on Instagram. Business Insider. Available at: https://www.businessinsider.com/deepfake-video-mark-zuckerberg-instagram-2019-6?r=US&IR=T (Accessed 24 Nov 2023).

European Parliament Research Service (EPRS). 2021. Tackling deepfakes in European policy. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf (accessed 3 Jan 2024).

European Parliament. 2023. EU AI Act: first regulation on artificial intelligence. Available at: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed 26 Jan 2024).

Hancock, J. T., Bailenson, J. N. 2021. The Social Impact of Deepfakes | Cyberpsychology, Behavior, and Social Networking. Available at: https://www.liebertpub.com/doi/10.1089/cyber.2021.29208.jth (accessed 13 Dec 2023).

Hansen, C. 2022. IBM. Available at: https://developer.ibm.com/articles/generative-adversarial-networks-explained/ (accessed 7 Jan 2024).

IBM. 2023. What are Neural Networks?. IBM. Available at: https://www.ibm.com/topics/neural-networks (accessed 7 Jan 2024).

IBM. 2024. What is Artificial Intelligence (AI)? Available at: https://www.ibm.com/topics/artificial-intelligence (accessed 12 Feb 2024).

Jaiman, A. 2022. Deepfakes aka Synthetic Media: Humanity at the Edge of an Uncanny Valley. Independently published.

Johnson, D., Johnson, A. 2023. What are deepfakes? How to spot fake AI audio and video. Business Insider. Available at: https://www.businessinsider.com/guides/tech/what-is-deepfake?r=US&IR=T (accessed 24 Nov 2023).

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Lei Ma and Liu, Y. 2022. Countering Malicious DeepFakes: Survey, Battleground, and Horizon. International Journal of Computer Vision. Available at: https://doi.org/10.1007/s11263-022-01606-8 (accessed 24 Nov 2023).

Khan, R., Bramah Hazela, Singh, S. and Asthana, P. 2023. Deepfakes: A New Era of Misinformation. ResearchGate. Available at: https://www.researchgate.net/publication/372058471_Deepfakes_A_New_Era_of_Misinformation (accessed 24 Nov 2023).

Lawson, A. 2023. A Look at Global Deepfake Regulation Approaches. RAI Institute. Available at: https://www.responsible.ai/post/a-look-at-global-deepfake-regulation-approaches (accessed 28 Dec 2023).

Lipsynthesis.com. 2024. Create Ultra Realistic Lip Syncing Videos. Available at: https://lipsynthesis.com/ (accessed 26 Jan 2024).

Mahmud, B. U., Sharmin, A. 2020. Deep Insights of Deepfake Technology : A Review. Available at: https://www.researchgate.net/publication/351300442_Deep_Insights_of_Deepfake_Technology_A_Review (accessed 30 Dec 2023).

Meikle, G. 2023. Deepfakes. Cambridge, UK. Polity Press.

Paris, B. and Donovan, J. 2019. Deepfakes and Cheap Fakes: The manipulation of audio and visual evidence. Available at: https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf (accessed 26 Jan 2024).

Paris, J. 2022. Deepfakes and the Uncanny Valley. Operational Excellence Society. Available at: https://opexsociety.org/body-of-knowledge/deepfakes-and-the-uncanny-valley/ (accessed 11 Feb 2024).

Reuters. 2020. Fact check: 'Drunk' Nancy Pelosi video is manipulated. Available at: https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated-idUSKCN24Z2BI/ (accessed 6 Feb 2024).

Serrels, M. 2020. Han Solo deepfake brings Harrison Ford to Solo: A Star Wars Story. CNET. Available at: https://www.cnet.com/culture/entertainment/han-solo-deepfake-brings-harrison-ford-to-solo-a-star-wars-story/ (accessed 15 Dec 2023).

Sloan, M. 2020. Deepfakes, explained. MIT Sloan. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained#:~:text=A%20deepfake%20refers%20to%20a,user%20of%20the%20same%20name (accessed 24 Nov 2023).

STT–AFP. 2023. ”Tytöt vain itkivät ja itkivät” – Tekoälyn luomat alastonkuvat ovat johtaneet skandaaleihin. Helsingin Sanomat. Available at: https://www.hs.fi/ulkomaat/art-2000010064263.html (accessed 12 Feb 2024).

Thompson, S. A., Maheshwari, S. 2023. 'A.I. Obama' and Fake Newscasters: How A.I. Audio Is Swarming TikTok. The New York Times. Available at: https://www.nytimes.com/2023/10/12/technology/tiktok-ai-generated-voices-disinformation.html (accessed 24 Nov 2023).

Timmerman, B., Mehta, P., Deb, P., Gallagher, K., Dolan-Gavitt, B., Garg, S., Greenstadt, R. 2023. Studying the online Deepfake Community. Journal of Online Trust and Safety. https://doi.org/10.54501/jots.v2i1.126 (accessed 3 Jan 2024).

Vincent, J. 2019. ThisPersonDoesNotExist.com uses AI to generate endless fake faces. The Verge. Available at: https://www.theverge.com/tldr/2019/2/15/18226005/ai-generated-fake-people-portraits-thispersondoesnotexist-stylegan (accessed 3 Jan 2024).

Westerlund, M. 2019. The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review. Available at: https://timreview.ca/article/1282 (accessed 16 Dec 2023).

Yasar, K. 2023. Neural Network. Enterprise AI. Available at: https://www.techtarget.com/searchenterpriseai/definition/neural-network (accessed 7 Jan 2024).

Appendices

Appendix 1 Request-for-input email sent to deepfake content creators

Hello …,

My name is Robert Schenk, and I am a student at Haaga-Helia University of Applied Sciences in Helsinki, currently working on my bachelor's thesis about deepfakes. I came across your youtube channel and noticed your familiarity with the concept and deepfake generation programs.

I am reaching out to you as part of my research to gain valuable insights from individuals with hands-on experience in deepfake technology. Your perspective would greatly contribute to the depth and authenticity of my study.

To express my gratitude for your time and insights, I am offering a reward of a $10 USD Amazon gift card or any other payment channel of your choice.

I kindly ask you to consider answering the following questions:

1. How difficult is it to create a credible deepfake?
2. In your opinion, is deepfake technology inherently dangerous to society? What possibilities, positive or negative, does it open up?
3. Deepfakes have been extensively covered in the media. From your experience, how much of the information surrounding deepfakes do you believe is legitimate concern versus sensationalized hype?
4. Additionally, I would appreciate it if you could take a moment to evaluate and provide feedback on a deepfake I created as part of this study. You can view the video at the following link: https://www.youtube.com/watch?v=GtFSXP9VpPI

Your expertise and insights will contribute significantly to the understanding of deepfake technology and its implications. Please respond by January 31, 2024, and I will ensure that the reward is promptly provided to you. Your contribution will remain anonymous. In the research paper, I will refer to you as "deepfake content creator".

Thank you very much for considering my request, and I look forward to hearing from you soon.

Best regards,

Robert Schenk

Haaga-Helia University of Applied Sciences

Appendix 2 Responses from deepfake content creators

Question 1 *How difficult is it to create a credible deepfake?*
Deepfake content creator no. 1 There are multiple open source programs created for deepfake creation - most commonly used DFL - which makes the process itself rather simple. It's just a data - computing power and time are the things most hobby users lack. Training a model in laptop GPU will take weeks and weeks for credible deepfake. To sum it up, it's not difficult at all, it takes time, money and lots of data.
Deepfake content creator no. 2 How difficult is it? Depends on how long you've been doing it. It requires moderate to high end PC hardware, so there's a bit of a prohibitive cost for some users. The learning curve is somewhat steep, but after you've done it a few times it gets fairly straight forward. After you've trained a model, you can re-use it on other material pretty quickly, so if you evaluate the difficulty at that stage, it's extremely easy (again, once you've mastered a few steps).
Deepfake content creator no. 3 It's not difficult as much as it is time consuming. You need to know where to gather the appropriate source data so the deepfake is as sharp and accurate as possible.
Deepfake content creator no. 4 Beginning from the ground up, creating deepfakes using DEEPFACELAB is a time-intensive process. It initially requires a couple of days to acquire movie rips and interviews, followed by extracting and refining the dataset. Subsequently, the model training phase can span anywhere from 1 to 7 days, with the duration varying based on the resolution and the GPU capabilities. However, leveraging existing datasets and pre-trained models can significantly reduce this timeframe, potentially bringing the entire process down to just 1-2 days.
Deepfake content creator no. 5 It’s more difficult than some might think. It took me quite a while to get the hang of it. I’ve only used DeepFaceLab which is supposed to be the easiest software to use but it’s still not exactly

easy. The face set you use is important and ready-made ones from the internet are not always good enough, so I usually make my own. Also, some videos are more difficult to deepfake because there is more than one face and you have to manually choose the face in each frame. Either way, there is a lot of work involved in the creation of a deepfake and you need a good computer either way to get any kind of convincing result.

Question 2

In your opinion, is deepfake technology inherently dangerous to society? What possibilities, positive or negative, does it open up?

Deepfake content creator no. 1

We recently did deepfake lip syncing example for CNN to educate public how upcoming American presidential elections will be manipulated. The article is not out yet, but I saw copywrite and their researched showed that malicious deepfakes have been used to scam 2.8 billion dollars around the world in 2023 only. More and more software's without set rules or ethics have been recently popping up and lack of regulations OR anti-deepfake software's from social media giants will pose giant risk(Intel claimed their anti-deepfake application can spot 97% of videos, we have not managed to test it yet, but I hardly doubt these claims are nowhere near for high quality videos). Risk of fraud videos/images have been in internet before, various celebs have been used for a decade to promote fraudulent health products and crypto scams using photoshop to manipulate images.

Currently, all we can do is educate general public to verify authenticity of the video that they see in social media, random websites or sent via message.

Deepfake content creator no. 2

Nothing is inherently dangerous. You can misuse an automobile, but most people own one. It's pretty much up to the individual to exercise judgement and behave responsibly.

Deepfake content creator no. 3

I think it can be dangerous to society when used politically or to damage someone's character. In my case I simply use it to change actors around in popular media with no malicious intent. If the subject of the fake were to ask me to take it down, I'd do it immediately.

Deepfake content creator no. 4

Like any technology, deepfakes hold potential for both beneficial and harmful applications. While I have never employed deepfakes for malicious purposes, the ease with which convincingly realistic videos can be produced raises concerns. The technology can be used positively, such as in aging or de-aging characters in films, resurrecting performances by deceased actors, enhancing remote
collaborations in the movie industry, using digital avatars for anonymity, or creating humorous content by superimposing celebrity faces. However, the darker side of deepfakes cannot be overlooked. Their use in creating inappropriate content, perpetrating scams, and fabricating political narratives poses a significant threat to society. For a majority of the public, approximately 80-90%, distinguishing between authentic content and these sophisticated fabrications remains a challenging task. This difficulty underscores the need for cautious and responsible use of deepfake technology.

Deepfake content creator no. 5

I think it is a risk to society. There are lots of creeps on the internet who are highly motivated to humiliate women by creating deepfakes of them, especially involving pornographic imagery. If you look at the available face sets online, a lot of them are of female gaming live streamers and other celebrities with toxic male fan communities. You can only imagine what kinds of deepfakes are being created of them. There are hardly any face sets of men. Coincidence? I think it's the most dangerous to private individuals. Having a deepfake like that created of you can be very damaging personally, even if you are not famous.

On the other hand, I don't think deepfakes are dangerous in terms of fake news. If someone creates a deepfake of Putin where he says he will nuke the entire world, I don't think that would cause any incident, no matter how convincing the deepfake is. People are quite sceptical of online media (at least I hope so).

Question 3

Deepfakes have been extensively covered in the media. From your experience, how much of the information surrounding deepfakes do you believe is legitimate concern versus sensationalized hype?

Deepfake content creator no. 1 You might say they are extensively covered, but are they? My grandmom is 75, I created her custom happy birthday wish from her favorite celebrity (Jeff Goodman, don't ask why she likes him). She started crying from happiness and there was not even a single second she doubted of the authenticity of the video. After saying this is NOT real, but AI generated, she said:" Oh really? I heard about this topic from TV but it looked so real." That kind of sums my point, even if something is covered, there will be still millions who will be fooled as the quality and authenticity get better over time and people are too lazy to run fact checks. It will take years and years before people will start doubting what they see and hear - overall, seeing should equal believing right? Not anymore ...
Deepfake content creator no. 2 I think the buzz around AI is sometimes misplaced, because people confuse one form of AI for another. Like people mention ChatGPT but then conflate that with Deepfakes or AI voices or Generative AI. I think the misuse of a person's likeness for let's say "NSFW" kind of content is an obvious concern. Election interference is another. Tricking people into believing a politician said something to cost them votes is pretty easily doable with the current level of technology, and it'll only get easier. On the other hand, you can use Deepfakes for pretty benign purposes like inserting your favorite celebrity into a movie they never starred in, or make your own content with them for fun without it being harmful to their reputation or having other negative outcomes.
Deepfake content creator no. 3 A lot of the information is overblown and most people who read or create concerned headlines about deepfakes don't understand the process behind them or do enough research.
Deepfake content creator no. 4 The deepfakes coverage in the media seem to be fair and legitimate imho. More awareness must be raised even at the cost of bad press.
Deepfake content creator no. 5 Deepfakes are a thought-provoking technology which sounds exciting and dangerous, but the deepfakes have been around for a few years now and most people are unaffected by them, positively or negatively. The risk surrounding deepfakes leaves a lot to the imagination. It is easy to imagine a scenario where deepfakes are used to influence global politics through targeted fake

statements by politicians. I think the people who make deepfakes and the people who talk about deepfakes the most are different people. Creators of deepfakes don't really draw attention to themselves, whereas news media loves to fearmonger and they are milking this for all its worth.

Question 4

Additionally, I would appreciate it if you could take a moment to evaluate and provide feedback on a deepfake I created as part of this study.

Deepfake content creator no. 1

I assume deepfacelab was used to create this. To create more realistic result, you need:
-Bigger faceset.
-Higher resolution model as original clip is HD quality.
-A LOT more training, depending on GPU, could be 3 days or 3 weeks.
-Different settings when merging(bad mask fitment, wrong colors, lightning is off).
-Proper post-editing - each time hand covers face, can see artifacts/glitches.

-Probably many more things that can be improved with practice. DFL is somewhat complicated software and there is no one workflow works method.

Deepfake content creator no. 2

If you want pure honesty, it's pretty clearly someone's early attempts at learning the process. I don't think enough source material was used to cover certain angles (like you can see sometimes when she should be looking slightly to the left, her face is instead slightly aligned to the right). Doesn't handle it well when she tilts her head back, her teeth are not well defined, and there's a bit of a transparency/opacity thing going on with her skin which makes me think maybe you used "seamless" mode instead of "overlay" when you merged it. A good first effort though.

Deepfake content creator no. 3

It's not bad for a start, but from what I can see the issues are resolution and overall dimensions of the encoders of the model. With a high resolution pre-trained model you can make a movie quality deepfake in about a day, if you also use a diverse and sharp faceset. However, with all deepfakes, it is very time consuming for the faceset preparation and training, but mostly the issue is it's expensive since you need a powerful GPU to create higher tier deepfakes. VRAM is always the bottleneck. Other than that, the other criticism is the face mask is a bit too small in

this example which causes the original face to show under, one would need to increase the size more and add blur to hide the seam.
Deepfake content creator no. 4 Pretty bad deepfake attempt that wouldn't fool anyone unless video got compressed way more.
Deepfake content creator no. 5 It’s a deepfake, but it’s not fooling anyone. It needs more training. Maybe a face set with more angles. The face gets blurry when she is looking even slightly away. Most importantly, the face needs to be blended better. I think with the proper blending options it would already improve a lot. Maybe make the video shorter. That way there is less material to process. So could be worse. Could be a lot better too.

Appendix 3 Frames from final deepfake

Figure 13. Frames from a deepfake of Anne Hathaway