Table of Contents

1. Introduction & Background

1.1 Relevance of the Topic

Artificial Intelligence cemented its presence and future potential in our daily lives recently. We wake up to a discussion on another billion-dollar AI deal of some tech giant almost daily. Infusing this technological transition into every sector one by one from beer production to medical analysis leaves us to witness a whole evolution before our eyes. It only seems uphill from here – if you wouldn't want to think otherwise.

The expansion from text to image creation, followed by video graphics, quite shook the tech industry and general public as an eye opener of AI's potential. It is possible to pick up your phone to be faced by a pseudo-Will Smith movie scene, and if it wasn't for the eerie distortion in every other frame, it would look entirely legit and plausible. As amusing as it is to watch, the emerging field of audiovisual AI raises controversies and challenges that require immediate action.

The main concern regarding audiovisual Artificial Intelligence is the autonomous use of copyrighted and private material during the creative process. A worldly amount of audio and visual data is processed and studied by AI algorithms which may easily induce concerns. The sky's the limit here; realistic deepfakes may unleash a whole new world in spreading misinformation and eruption in news media.

Fortunately, there have been recent developments in establishing a legal landscape in the field which this thesis plans to dive into, with the **EU AI Act: the first regulation on artificial intelligence** and its takes on the audiovisual aspects being the primary focus. By deeply exploring the areas of copyright infringement, privacy precautions, and the proposed legal frameworks as potential

solutions, this thesis aims to contribute to the deployment of the future of mass media and communication technologies.

To get a good grasp of the complex nature of audiovisual artificial intelligence, the approach within the following chapters is going to be evaluating its respective individual components separately. Examining the unique insights and natures of audio and visual AI in isolation will display their distinct challenges and specific ethical and legal challenges more clearly before building up a good foundation in understanding the future discussion of the combined challenge, that is audiovisual AI with its broader implications.

1.2 Morphing the Landscape: Audio vs. Visual AI

1.2.1 Audio AI

Audial artificial intelligence is revolutionizing every method where technology interacts with sound with its special focus on audio data manipulation and processing, primarily impacting the technologies in communication, information, and music. Being a spectrum that ranges from keyword-prompted music generation to multi language voice assistants, the diverse applications integrate a natural element into the analytical processes to make human interaction with the machine creative and intuitive, almost as if we are communicating with a friend.

Sympathizing the user in interacting with the data-hungry algorithms that power most of the Audio AI tools to eventually overshare personal information which might, in fact, be one data breach away from revealing access to the user’s main daily life and financial components raises privacy concerns across the subfield

and reveals disadvantages aside from the positive aspects, both of which are explored in the following sections.

1.2.2 Applications of Audio AI

The most prominent aspect would be its use in the music field without any doubt. Recent algorithms are known to personalize playlists to adapt to user preferences. For instance, Spotify has adopted an algorithm to craft its main playlists depending on user preferences. From “Songs to Sing in the Shower” to “I Love My 90’s R&B,” the platform aims to study our music taste profile to shed it seamlessly across popular playlists to make the overall app experience *feel like home*. Spotify provides an insight into the process: as we delve hours into automatically curated playlists without a care, the machine learning algorithms in the background – with this one named “Algotorial” – study our listening history and preferences, and match tracks from the editor-created “song pools” dedicated for each type of playlist depending on our favored rhythm, melody, and genre by analyzing audio throughout the whole process.

The streaming platform did not stop there. Spotify is now further implementing artificial intelligence into its practices by introducing a tool to create personalized playlists in return for text prompts: users can simply input a description of a theme, occasion, or mood for the algorithm to analyze their listening history with respect to it and generate the desired playlist, which will remain available for further modification. Retrospectively this development might sound “just trendy;” however, it takes a second of comparison to the noughties, when playlists would be specifically hand-picked and burned to physical CDs, to reflect how far the field has developed.

Next, AI-powered voice assistants have become an essential life companion overnight. From homes, refrigerators, and cars to mobile devices, these assistants rely on natural language processing and speech recognition, thus able to comfort even senior consumers to familiarize themselves with the technology to build up long-term trust, value, and loyalty.

Furthermore, this subsection of artificial intelligence has an important role in offering accessibility for individuals with disabilities: transcription tools run by AI provide speech-to-text services for the visually impaired, and real-time media captioning tools widen information access and help the digital environment become more inclusive. And not just for the ones in need, these developments also help break language barriers by offering AI-powered real-time speech translation with voice recognition. Hence, today it is possible to witness a worldwide summit where every speaker effortlessly communicates using their native languages.

While we are able to cover much of the benefits of this subset, applications surrounding Audio AI also come with their challenges that require thorough attention and consideration, with the foremost example being the viral AI-generated music content that is spread throughout social media. It is an inevitable and irresistible source of entertainment: when done beautifully, it is impossible not to adore and appreciate the technology in awe. Let it be a current artist covering an old song, a late artist covering today's hits, a younger version of the artist covering their own song, or vice versa – the possibilities are endless.

Here's an example. Gaining over two million views, it is truly a smooth work that displays how successful the technology is. However, the work is not the

real shocker: the YouTube video, uploaded by what seems to be a regular user, is in fact copyrighted and monetized by the original songholder. This means that the credit goes to Adele, even though Lana Del Rey is technically the source of the voice in the content. Thus, a brand new type of intellectual property taboo arises within the advancing field of artificial intelligence. Who owns the rights: the training data, the programmer of the AI software, the user of the software, or the voice owner?

A further study of this application appears in a similar product that is deepfakes. Being a malicious tool of misuse for disinformation, deepfakes can go as far as to damage reputations and manipulate masses through the media. For instance, imagine the spread of false statements by a renowned politician during an election period. It would pose a critical threat to the trustability of the targeted party and media services, as well as harm to public discourse's integrity.

1.2.3 Visual AI

Visual artificial intelligence makes up the latter part of the discussion as it is a literal "eye-opener" for computers and machines by introducing processing capabilities of the visual environment. Daily devices have been long using and advancing facial recognition technologies from identity verifications to personalized marketing. The technology that enables all the visual specializations in the field is object detection, which teaches AI to autonomously identify and label objects within visual media. It plays a giant role in various sectors nowadays, ranging from self-driving cars to medical analysis. Furthermore, visual AI also transforms the media editing processes by offering auto-editing, background/object removal, as well as from-scratch generation of realistic images and videos.

Overall, much like its former companion audio AI, visual AI also transforms everyday processes to redefine the way humanity creates and consumes media, with several highlights explored in the following sections.

1.2.4 Applications of Visual AI

The benefits to the visual landscape introduced by AI are vast and they strongly enhance user experience. The foremost example can be improved visual search, which uses help from AI technology which understands and studies image and video context to provide more accurate and efficient search results according to users' search queries. Think of a funny video you refer to your friends only to be met with confusion and blank stares. Then you go over to the internet to show your reference; however, you do not know how to label the content since you did not search for it in the first place - you just happened to come across it one day. Luckily, you can name or describe an object in the video and you are most likely going to succeed in finding the content even if the video is titled something completely different. Next, the content creation process today is just as amusing. AI can now automate or initiate tasks in the video editing process, freeing content creators to be able to focus more on the creative process of production. From object frame focus to complete object removal, possibilities are only expanding.

Finally, visual AI bridges the gap for the visually impaired by a real-time description of visual content to assist them in navigating their surroundings and getting a good grasp of their environment. Developments especially like this one truly shed a bright outlook on the future of the technology field and leave one hoping to witness more advancements as soon as possible.

Although developed for the good, much like the audio field, there is an abundance of data collected and studied by the companies specializing in the topic for the process. Acknowledging the potential risks and controversies regarding this data is crucial as a consumer and data subject.

At the foremost, a significant concern arises in privacy. Our faces are facing a smart device almost all day, and the facial recognition technology - or even a simple front-facing camera - that's built-in in these devices deceives our interactions and activity down to the tiniest bit. An overnight data breach or unauthorized access and study done by some tech giant could simply erode privacy and technological perception. It is in fact more vulnerable than one would guess: a recent study done by 3M concluded that visual data breaches are an upcoming concern within the AI security field, given that some 67% of employees share that they work with confidential data outside their offices, preferably in public places, and 70% of companies are insufficient in their policies about it. This creates a critical threat to the information to be monitored and stolen from devices even with privacy screen protectors, which have been found to be not quite useful for the purpose. This surely creates a need for thorough visual AI security measures and an overhaul by the company for their coverage on the topic.

Another worrying implication is no other than the manipulated media called deepfakes, which leads the discussion throughout this paper. In collaboration with Audio AI manipulation, the creation of realistic synthetic media can be used to promote the spread of misinformation in order to damage public perception of targeted persons or ideas. Aside from the politician example from section **1.2.3**, another harmful use can be the risk of manipulating image recognition algorithms to bias them towards favoring a certain race in recruitments due to training data.

Overall, if we focus on responsible development under ethics and regulations, Visual AI can be a powerful tool that's for the good of society. As much as the technical development, navigating Artificial Intelligence development through moral barriers and norms is also part of its core development and consolidation.

1.3 Rise of Deepfakes, Need for Regulation

We are going through a time where our daily lives and gadgets are prioritized to intertwine with AI in every process possible, and the branch that focuses on audiovisual aspects especially shines with never-seen-before methods: movie/media recommendations personalized for the user on streaming services, voice assistants, music & image generation, and many more which have already been covered throughout section **1.2**.

The benefits yielded by the technology are inspiring. One can design a future where education has become personalized and adaptable to everyone's learning style, pace, and preference. It is not far from reality: AI can offer the experience by analyzing worldwide student data and designing the appropriate content and algorithms. A current study on the field is done by Google with its new generative AI project called LearnLM, aimed to make education more "active, personal, and engaging." Furthermore, numerous educational institutions are slowly integrating AI solutions after the recent popularity of chatbots among students. The education integration will reshape the problem-solution approaches run by AI that the new generations will familiarize with, ranging from fixing backend coding to music composition.

Nevertheless, behind the exciting evolution of these advancements in the audiovisual AI field lies a deepening concern: deepfakes. Deepfakes are AI-generated synthetic media that is able to manipulate visual or audio content in order to create ultra-realistic imitations. Take a look at this example: it is an equally impressive and scary one. If one was to come across the video in a mindless scrolling session, they would think it is literally Donald Trump himself. Apart from the occasional graphical stutters, the video is a perfect display of what an individual with regular technology access is able to do.

Deepfakes achieve its purpose by leveraging complex Artificial Intelligence techniques such as Generative Adversarial Networks (GANs). GANs can be considered as a system composed of two AI algorithms in competition with one creating the imitation content, called **the forger**, and the other trying to identify the counterfeit, called **the detective**. During the process, the forger constantly learns and improves the creation of the deepfake while the detective augments its ability to spot and monitor the content. The backend of the application gets increasingly complex, however it is essential to understand what lies at the core of deepfake creations. The details of GANs are further discussed throughout **Chapter 3**.

Overall, deepfake creation is spreading in popularity, accessibility and versatility by each day, despite not being necessarily always for the welfare of all. The superficially innocent outlook and potential of the tool takes a sinister turn when its malicious applications are explored, and these state-of-the-art forgeries are a growing threat in criminal and unethical activities.

A major concern arises within the use of the deepfake technology for impersonation fraud. Consider a video meeting for work with your CEO and other

team members, all elaborating on the urgency of a giant financial transaction. Though, in fact, the other meeting members are not human; they are deepfakes! This became the reality of a finance worker from Hong Kong back in February. He was ultimately believed into paying out some $25 million to disguised fraudsters through a fabricated video conference. Despite initially being suspicious by a phishing mail, he let his doubts die down after the successful deepfake performance during the meeting. The situation was realized after the Hong Kong police revealed the AI involvement in the incident, when the damage was already done.

The spread of misinformation is another case where the technology wreaks havoc. The growing presence of deepfakes in the political scene comes at the foremost of concerns. Several headlines of AI incidents have taken place in the US and India during their electoral runs. For instance, a manipulated Joe Biden voice was presented to Democrats to discourage them from participating in the primary elections, and India set the sky as the limit with their deepfake usage of the candidates, going as far as holding pretend TikTok live streams.

These are just several instances headlines are starting to familiarize with. As the technology advances even further, such events have the potential to cause a dramatic impact on public perception and behavior. Therefore, this raises the urgency for an expansive regulation and mass education. Otherwise, digital media is vulnerable to become a blurred line between fiction and reality, creating an environment where trust is eroded and consequences are beyond imagination.

A robust legal framework is a priority in the digital landscape today if authorities aim to mitigate risks. While there are current legal frameworks which mainly address intellectual property and data privacy issues, they may appear

outdated and lacking at the deepfake technology scope. The gap between existing legal frameworks and the rapid technological breakthrough arises an ethical vulnerability that may be exploited.

To answer this rising concern, the European Union (EU) took a proactive approach by proposing the AI Act. This milestone legislation focuses on regulating the development, implementation, and use of Artificial Intelligence in a wide range of aspects, affecting both citizens and authorities. Especially focusing on high-risk applications, the EU AI Act raises a crucial question: can it successfully address the challenges related to deepfakes?

This question will form the core throughout this thesis. By evaluating the provisions of the AI Act and pointing out its limitations in deepfake regulation, as well as support by several current studies and proposals done on the field, we will asses the effectiveness of the bloc in combating such an emerging threat. Additionally, by focusing on areas with room for improvement and in need of more comprehensive and forward-looking legal frameworks, it will be possible to pave a future where the deepfake technology is practiced ethically and responsibly.

1.4 The EU AI Act: A Sufficient Response to Risk?

As discussed, The European Union (EU) pioneered to advocate an answer to the growing concern about the misuse of AI with a responsible deployment framework. The EU AI Act, a recently adopted product of this intent, is the key piece designed about the use of new AI technologies across sectors.

The AI Act distinguishes AI applications across various risk levels, going from loose to more strict regulations as the application comes closer to

“high-risk.” Such applications undergo thorough examination including human oversight mechanisms, data governance programs, and mandatory overall risk-calculating assessments.

One of the forthcoming aims of the AI Act is to mitigate the risk of manipulation with Artificial Intelligence. Specifically mentioning “deepfakes” as a concern in the category, the Act targets the manipulative and destructive outlook of the tool. As a result, several approaches are constructed. Firstly, the Act targets increasing transparency in the field by demanding a clear labeling of created media content. The creators are asked to disclose the techniques and motivation of use to help users be able to distinguish between genuine or fabricated media. Then, the deepfake technology developers are asked to hold rigorous risk assessments to legally identify & mitigate potential dangers in their practices, which will be monitored by authorities. Finally, the Act highlights the importance of human observation in high-risk applications to include mechanisms that would prevent misuse of deepfakes for illicit purposes.

While the EU AI Act promotes a step towards regulation of deepfakes, we are yet to acknowledge how effective it is in application. Therefore, the following chapter will dwell deeper into how the Act plans to tackle the challenges that come from deepfake content by analyzing the potential impact with its strengths and limitations.

1.5 Thesis Statement

Although the EU AI Act recognizes the manipulation risks within AI and therefore initiates steps to target deepfakes, this thesis approaches the Act from a

perspective that questions and evaluates its sufficiency. The current framework might contain insufficiencies in adaptability and scalability. By critically approaching the provisions and limitations in the regulation of the technology, the paper explores the following question:

Is the EU AI Act an effective answer to the emerging threat of deepfakes and its misapplications?

The evaluation will include assessing the strengths and weaknesses of the Act, exploring potential loopholes, and raising the necessity for a more comprehensive legal framework that ensures ethical and responsible use and development of the deepfake technology.

2. Legal & Ethical Considerations

2.1 Importance of Ethics and Regulation in AI & Deepfakes

As discussed in the first chapter, swift advancements in Artificial Intelligence have unleashed a new world in technology. Nevertheless, its undeniable benefits demand a responsibility to ensure an ethical use and development of its tools. More specifically, AI ethics is an important field that emphasizes the moral implications of the technology. It questions the transparency, fairness, and accountability of the structure, management, and impact of Artificial Intelligence systems.

Why do we consider AI ethics to start with? Lack of regulation in the field leads to a range of negative consequences, with the foremost being bias. The algorithms depend on their training data for performance and wellbeing. If the training data consists of inherent bias, the end product can conserve these in its decision-making process. This could lead to discriminating results in areas such as job applications or loan approvals.

Organizations have recently discovered several high-profile bias examples within AI tools in a wide range of fields. First example is in the healthcare field: computer-aided diagnosis (CAD) programs are discovered to output lower accuracy for black patients compared to white patients in its results. Next there is the advertising field, where a research in Carnegie Mellon University in Pittsburgh has discovered that the online advertising system of Google displayed higher paying positions to males more often in comparison to women. Furthermore, a generative AI art generator called Midjourney was found to display men noticeably more often than women when asked to generate people in specialized positions. Lastly, predictive

policing tools that are run by AI are used by several organizations in the criminal justice system to identify locations with higher probability of crime. Relying on historical data, the prediction tools are influenced to apply the existing racial profiling pattern to disproportionately target communities of the minority.

Another concern in AI ethics is privacy. To perform well, Artificial Intelligence systems need to access an abundance of personal data to train on to begin with. However, if appropriate safeguard measures are not taken, the collected training data might possess serious concerns of privacy. Individuals might subconsciously be profiled, targeted, tracked, and manipulated by AI algorithms that are leveraging their sensitive information.

These were just several standouts of the potential pitfalls of lack of regulation in the AI field. If the necessary ethical regulations and guidelines are established, there can be a solid progress towards ensuring that the AI technology is mutually beneficial for developers and consumers. Transparency, fairness, and human rights will be satisfied, which will further solidify the trust of consumers.

Nevertheless, the case of deepfakes need special care and attention. Ethical considerations, even though an essential foundation for responsibility in development, will not be enough by itself to cover a wider scope of Artificial Intelligence. Unlike the other tools, deepfakes showcase the ability to manipulate human perception. Even if bias and security needs are satisfied, deepfakes can train on legally safe content amidst aiming for a sinister output. How can we make sure that this technology is practiced responsibly?

As dwelled throughout sections **1.2.6** and **1.3**, the threats posed by deepfakes to society include erosion of democracy and free press, going as far to

become a weapon aimed to destroy reputations as discussed with several examples. The current ease of reachability and distribution of the tool urges imminent legal frameworks to complement ethical considerations alone. Liability, accountability, and transparency should serve as the basis for the solution.

Additionally, deepfakes raise the need for international cooperation. By combining robust legal frameworks with all ethical principles, as well as maintaining a constructive dialogue and collaboration within technologists, programmers, policymakers, and the civil society, deepfakes can eventually be reintroduced to serve a positive purpose.

2.2 Limitations in Data Protection Laws

Data protection laws endure an essential role in safeguarding sensitive information of citizens, such as the General Data Protection Regulation (GDPR) law throughout Europe. Such regulations enable individuals to be able to make decisions over their private data, letting them choose how it should be collected, stored, and used. Furthermore, these data protection laws focus on diminishing harmful practice in favor of more responsible applications by demanding liability and transparency from controllers and managers.

While laws such as the GDPR are essential, they do possess limitations in regulation of AI subjects including deepfakes. For instance, there is a challenge in classification. Deepfakes blur the lines of what parts of its content constitute a violation of personal data protection with its nature. A deepfake video might be generated by studying visuals that are of public domain. This would not

directly infer an unauthorized use nor theft of personal data. Such a scenario might strain the data protection law to defend a legal basis to intervene.

Furthermore, it is a whole separate challenge to identify the minds behind deepfakes. They often anonymously operate or disguise themselves as ambiguous online identification. This encryption introduces new barriers for authorities in data protection to identify the accountable creators and enforce the necessary regulations for their breaches.

There is ongoing preparation, discussion, and proposition to extend the scope of laws in data protection, including the GDPR, for improving the mission of addressing deepfakes. An idea is to implement deepfakes as a standalone class among the concerns for data protection, even in the case where sensitive personal data is not directly stolen. In addition, the minds behind deepfakes can be asked for a transparent disclosure to explain the purpose of using the manipulated media, therefore allowing consumers to make more informed decisions on the information that might encounter them.

These are just two of the many avenues that invite data protection laws to expand and improve on for the mission of effectively tackling illicit workaround in deepfakes. Overall, the legal field raises a complex issue that requires thorough study and evaluation to perfectly balance the needs of freedom of expression and regulation.

2.3 The EU AI Act: the Current Answer

Following the advocacy of a much needed robust framework for Artificial Intelligence, the European Union pioneered in establishing a set of

regulations to reign over the technology. Passing in March of 2024, the EU AI Act showcases a groundbreaking body of work that strikes to balance technological innovation with civil safety and ethical considerations. It consists of a framework that classifies Artificial Intelligence systems based on their respective risk levels, with the regulations proportionally increasing with risk. The Act also puts special emphasis on deepfakes, which is our main motivation of evaluation throughout this section.

The Act calls to prevent spread of disinformation, safeguard democracy, and protect fundamental human rights while also endorsing AI development. As introduced, The Act establishes three risk levels for AI systems: low-risk, high-risk, and unacceptable (Article 3). Throughout Article 6, high-risk applications with stringent requirements are detected in the following areas:

1. AI integration used in facial recognition and other similar biometric identification practices.
2. AI systems deployed throughout sectors including energy, water management, and transportation.
3. AI systems that influence professional, academical, and educational trajectories.
4. AI integration in professional recruitment, employee monitoring and worker management.
5. AI implementation in essential public and private services, such as healthcare and utilities.
6. AI practices in law enforcement and criminal justice activities.
7. AI systems in managing border control, immigration, and asylum.
8. AI practices that impact democratic activity and court actions.

It is strictly noted throughout Articles **10-15** that such high-risk applications must fall in line within set requirements on data governing, technical documentation, repositories, human monitoring, accuracy, robustness, and security.

The AI Act is set to mitigate manipulation risks, the key aspect of deepfakes, within various measures as the following:

1. **(Article 10)** Developers must offer a risk management mechanism to automate the process of identifying, analyzing, and mitigating the risks in their AI practices, aiming to minimize the unethical practice of deepfakes.
2. **(Article 13)** The developers of AI systems who partake in manipulated media generation must particularly communicate the nature of the content, possibly by the use of labeling and disclosure on their model's capability & limitations.
3. **(Article 13)** AI models must embed solutions for marking content in a machine-recognizable format. Metadata detection, cryptography, logging, fingerprints, and watermark techniques should make sure that the content is validated as AI-generated.
4. **(Article 14)** High-risk applications must contain human-detectable mechanisms to avoid risks of manipulation or misperception.

Aside from the challenge of monitoring, the presence of deepfake applications alone causes several obstacles. Firstly, The AI Act does not directly disclose deepfakes as high-risk. The classification of an application as high-risk is more dependent on the application context and not directly about the technology, which leads to ambiguity. Then, the deepfakes purposed for entertainment or other naive purposes might not make the cut as high-risk despite its evident potential for fraud or misinformation.

Following the discussion, the set of approaches of The Act to address deepfakes are not clearly defined, therefore allowing for potential workarounds due to regulatory gaps. There are no specific guidelines set to enforce rules against the misuse of deepfakes for overseeing authorities (Article 45). While there are penalties applied for non-compliance within a broader context (Article 44), there are no direct deepfake-specific sanctions enclosed. Consequently, user accountability and content removal becomes a problem. There is no current robust framework that suggests an immediate takedown of malicious deepfake media, which would be effective in mitigating potential damage. There is no specific focus on holding deepfake creators liable for any malpractice either. The Act does target Artificial Intelligence developers (Article 10), however there is a gap in targeting individual accountability. Such lack of focus on liability inevitably weakens the ability to deter creators to incline towards mispractice.

Artificial Intelligence is currently experiencing new peaks of growth, and it is rather acceptable that the extensive framework of the EU AI Act may need additional adjustment to catch up with AI advancements. Therefore, several insufficiencies of the EU AI Act within the target of deepfakes surely come with potential solutions. The Act could be revised to also hold deepfake creators individually accountable along with AI developers to maximize efficiency in mitigation. Moreover, monitoring of media organs and social media platforms could be expanded to include real-time reporting algorithms and temporary suspensions for users consistently sharing misused deepfakes. By addressing such limitations. The Act can become more scalable with a more robust framework for regulation. Throughout the following section, these potential solutions and current discussions will be more thoroughly explored to assess the effectiveness of the AI Act.

2.4 Limitations of the EU AI Act on Deepfakes: Deepening the Discussion

As covered throughout the previous section, the EU AI Act poses several shortcomings in regulation despite pioneering a significant step in the field. Now we dig deeper into such limitations by exploring obstacles in mechanisms, content monitoring, and classification, highlighted throughout Recital 79.

As discussed, the Act consists of risk categories where the most stringent regulations are imposed on high-risk systems. Though, shaping these regulations for the case of deepfakes is not a perfect solution. This is because the AI Act does not define “deepfake” directly, and therefore does not categorize them as high-risk either. Due to the technical-based context within classification, the ambiguity allows for loopholes and uncertainty. As a result, it should be a priority for the EU commision to provide a more direct approach to deepfakes within the AI Act. A potential sub-category within the high-risk would satisfy the needed scalability.

Moreover, despite the established framework that’s suited for deepfake regulation throughout Articles 44-45, the enforcement mechanisms are still not clear. The Act discusses that the national authorities of each member state should maintain the oversight, however there is potential for inconsistency between members due to the ambiguity of guidelines for deepfakes. Plus, we already discussed in the previous section about the proposed penalties in case of non-compliance not being specific and deterrent enough. As a consequence, the set of standards for each member to follow should be more clearly explained, and the sanctions should be strengthened.

Lastly, we discussed the insufficiencies in targeting imminent monitoring and dismissal of malicious deepfake media. By satisfying the proposed real-time tracking systems for removal and creator liability, The Act will satisfy the fundamental and strong requirements to become the up-to-date and sufficient answer to the emerging field of Artificial Intelligence we face today.

3. Deepfake Deep Dive: Technological Methods & Legal Implications

3.1 Exploring Deepfake Development Techniques

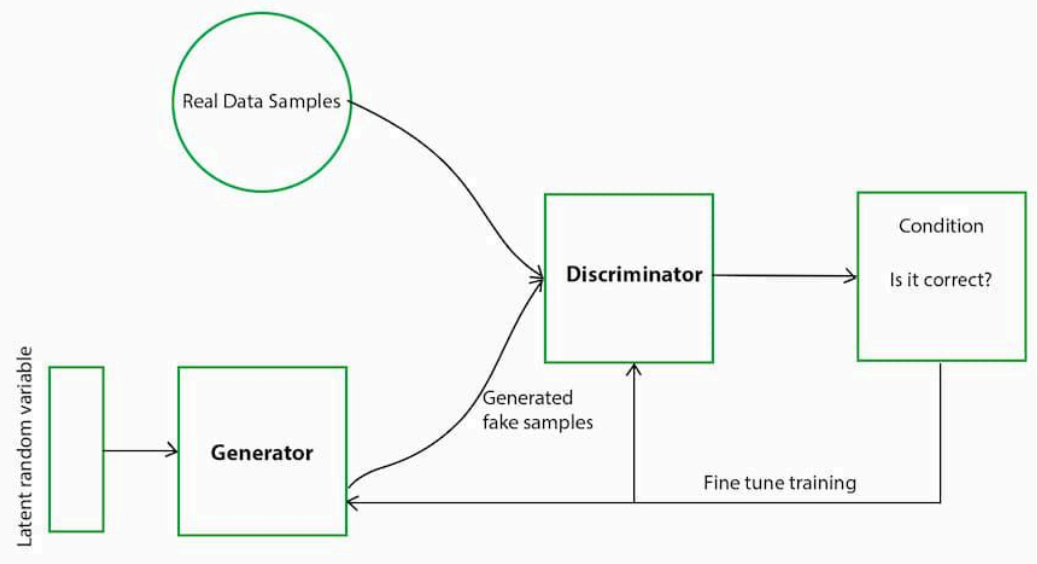
3.1.1 The Generative Adversarial Network

As briefly introduced in section **1.3**, the deepfake technology relies on the subfield of AI that's called deep learning in order to generate real-like synthetic media. These models belong in generative AI as a whole, which includes GANs, Transformers, Variational Autoencoders, and Diffusion models. The models are roughly influenced by the human brain, and are studied on extensive audiovisual datasets. With such training, the algorithms leverage on complex patterns and relationships in the data. Then, on the deepfake side, the deep learning algorithms analyze the source visual data to come up with new pieces of realistic synthetic media to simulate realistic but fake scenes.

Today, the most advanced model to create deepfakes is the Generative Adversarial Network, shortly called GAN. The GAN network is made up of a pair of neural networks: the generator and the discriminator. The generator is deployed to generate the faked media from scratch - for instance a false news report. Then, the discriminator analyzes the generator's output to distinguish it from the actual source data, therefore acting as a critic. The discriminator is also usually implemented as a Convolutional Neural Network (CNN) for image data.

Then, both components work in iteration and opposition within this order during adversarial training. The generator keeps submitting its creations to the discriminator and then refines it according to feedback until the point where the generated media becomes realistic enough for the discriminator to not be able to distinguish from the source data. It is nicknamed a "minimax game," since the

generator attempts to fool the discriminator while the discriminator leverages its identification capabilities. Interestingly, when the generator improves by creating more realistic data, the discriminator also advances by identifying these better fakes. Overall, the accuracy of the model output is optimized when the discriminator cannot reliably distinguish between real and fake data. Overall, the goal is to minimize both networks' loss functions for the best possible performance. Here is a visual breakdown of the GAN working architecture and the flow of information:



Source: GeeksforGeeks

As a result of the iterative adversarial training process, GANs are able to generate convincing products of deepfakes within limited data. Also learning and flourishing its model during each creation process, GANs are currently able to mimic voices, faces, and entire bodies with movements and dramatically reduce the amount of necessary training data.

Additionally, Multimodal AI is capable of noticeably enhancing the application and functionality of GANs. It can integrate various data types including text, audio, and images, allowing GANs to surge in accuracy. For instance, a

multimodal GAN can generate an output with synchronized audio which ensures that the generated speech is in sync with the facial expression and lip movements of the speaker. Overall, this combination improves the coherence and believability of the generated output, making multimodal AI a notable tool in training.

While the GAN tool is powerful, deepfakes leverage several other alternative approaches as well. First, facial recognition technology, as discussed throughout the first chapter, has an essential role in data collection and utilization for deepfakes. It studies the key features of the human face such as mouth, nose, and eye positions. Then, such collected information can be leveraged by AI to generate deepfakes with even more human-like movements and expressions. Then there's the voice cloning technique, which involves synthesization of new speech using voice recordings. There are multiple available approaches to obtain this, including deep learning or statistical parametric systems. Ultimately, the result of voice cloning will serve as the audio source for a successful deepfake project to target, say, a renowned figure.

3.1.2 The Celeb-DF Algorithm

Highly advanced datasets for deepfake training are in fact quite accessible online. As an example, we have the Celeb-DF dataset, consisting of 5,639 advanced deepfake celebrity media. The dataset shines through its competitors in the way that it does not suffer from low-quality or commonly seen source visuals thanks to its specialized synthesis methods in training. Here is a more detailed comparison of similar datasets:

Dataset	# Real		# DeepFake		Release Date
	Video	Frame	Video	Frame	
UADFV	49	17.3k	49	17.3k	2018.11
DF-TIMIT-LQ	320*	34.0k	320	34.0k	2018.12
DF-TIMIT-HQ			320	34.0k	
FF-DF	1,000	509.9k	1,000	509.9k	2019.01
DFD	363	315.4k	3,068	2,242.7k	2019.09
DFDC	1,131	488.4k	4,113	1,783.3k	2019.10
Celeb-DF	590	225.4k	**5,639**	2,116.8k	2019.11

Figure 1: Information on Various Deepfake Datasets

The table details the numbers of real vs. deepfaked videos and frames for each notable dataset including how up-to-date they are, and Celeb-DF clearly leads the way with the largest scale of video collection. Therefore, it serves as the go-to choice to replicate celebrities for any purpose one can imagine, and beyond.

Then, the second figure shows the common insufficiencies seen in deepfakes across different datasets. The lack of quality, color and positioning mismatches, splicing boundaries, and orientation inconsistencies are clearly visible which highlight the results from the previous table, showing how less data in recency and scale result in poorer results.

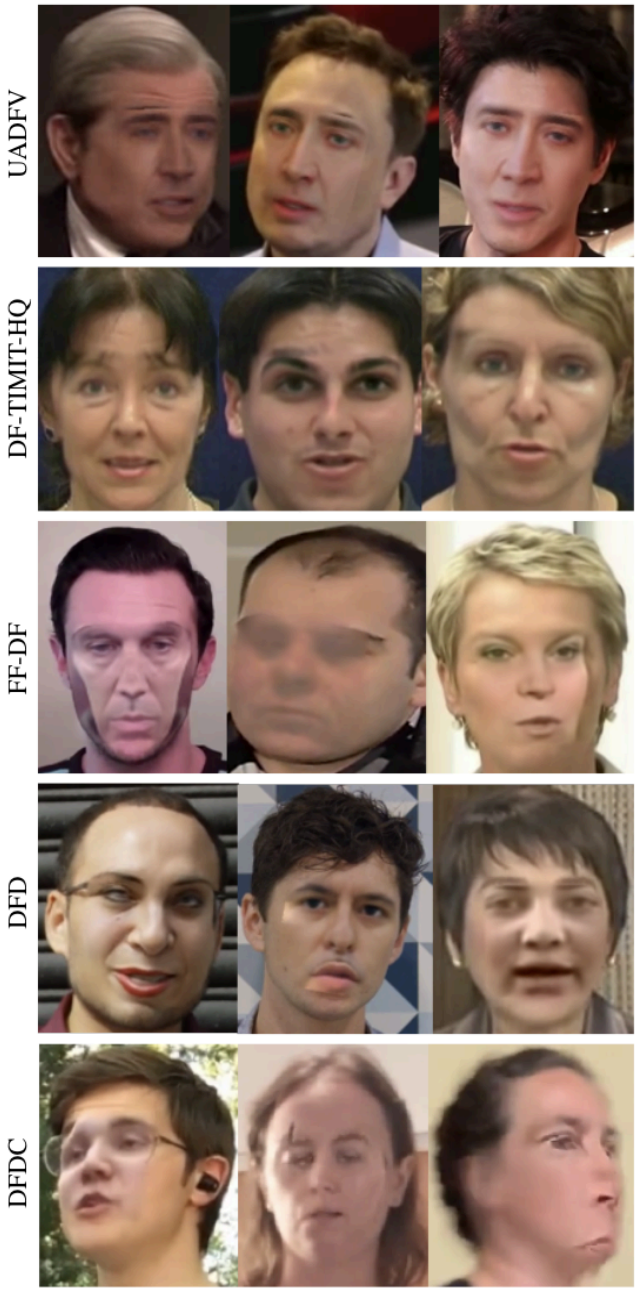


Figure 2: Visual Artifacts from Poorer Datasets

Up next, this following figure details the process of training and synthesis of the specialized deepfake algorithm for Celeb-DF. While the left part displays the synthesis process, the right side showcases the training side that uses the help of an auto-encoder to capture the targeted face to replace it with its synthesized one. The technology ensures that the pose and expressions remain for

convincibility. The continuous advancement and enrichment of this algorithm and the source dataset is key for Celeb-DF to ensure differentiation from its peers.

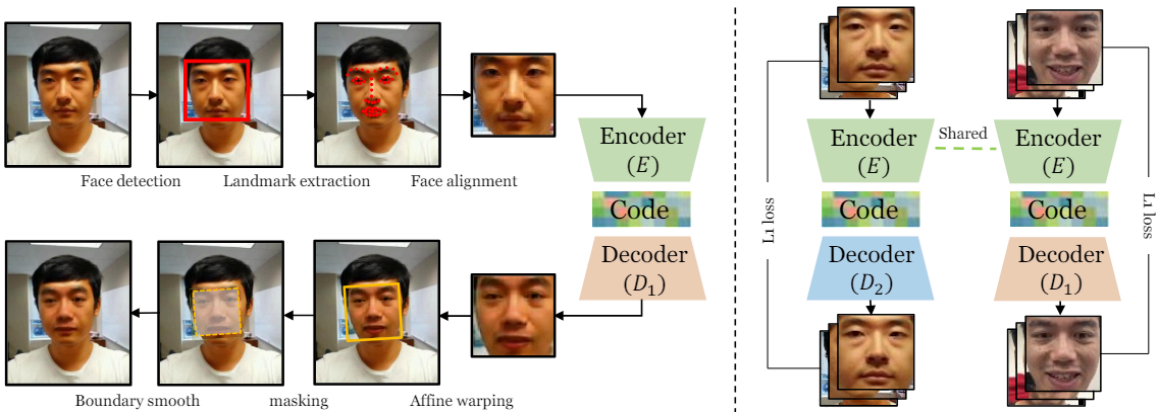


Figure 3: Celeb-DF Algorithm

Lastly, the following figure presents the Celeb-DF generated frames, showcasing its quality and differentiation. The real video footage sits at the leftmost column, while the other five display the deepfaked frame variations. Through all four rows, the subjects used are mainly renown celebrities including Jim Carrey and Brad Pitt. Overall, the figure perfectly visualizes the quality and realism of the current state of this technology. The technology seems to be ready; at this point, it is left to the developers; creativity as to what kind of content they would like to offer.

Figure 3: Frames Resulted from Celeb-DF

As for the legal side, The AI Act should ensure to satisfy the much needed focus on finding an answer to all these advancements in GANs, deep learning, and similar alternative technologies which promote deepfake development. Artificial Intelligence currently embodies an ever-evolving nature. As discussed, the regulatory gap caused by this could cause confusion within the implementation of the Act with its definition of “high-risk” AI applications as of now. Therefore, adaptability and flexibility should also be prioritized in a possible revision of the EU AI Act framework.

3.2 Deepfakes: Spectrum of Uses

As covered with many examples throughout the previous sections, deepfakes have a vast range of possible uses, both positively and negatively. Acknowledging and understanding the areas of practices as much as possible is key to efficiently assessing consequences of the technology and regulation. This section will recite these common applications collectively.

Formerly, as discussed in the first chapter, the positive applications of deepfakes primarily include special movie effects. The enhancement offered by the AI technology is revolutionary compared to what manual editing can achieve in the same amount of time. Deepfake technology in this field offers body doubles, de-aging, and even cameos by deceased actors while also cutting down on time and cost.

Next, deepfakes are being implemented into education. With the help of AI and analytics, deepfakes are utilized to create personalized content for education. For instance, deepfaked voices may be implemented in language practice tools for more immersive and engaging sessions. The ability to interact with mimicked native accents and pronunciations powered by AI can for sure help in personalizing the education further compared to traditional methods.

Furthermore, deepfake technology offers a breakthrough positive development in accessibility tools. For example, people with speech impairments can use the help of deepfake tools to create synthetic speech that mimic their own voice. This will offer a more personalized communication aid for the ones in need.

Followed by the positive applications come the negative areas of use. At the foremost comes the creation of non-consensual pornographic content, specifically involving celebrities. Such content is created to monetize on, ignoring the reputational and emotional damage of the subject. Recent headlines have surfaced in January when the famous singer/songwriter Taylor Swift became the victim of deepfake pornography, gaining over 45 million views on X. The incident received a response by European authorities to acknowledge the criminal activity of such content, with hopes to protect citizens from the abuse. The proposed legislation is due 2027 as part of a broader framework targeted towards cyberviolence.

Then, as highlighted throughout sections **1.2.3** and **1.3**, the weaponization of deepfakes within the political field is becoming more and more common nowadays. In broad terms, the use of deepfakes for unauthorized impersonation fraud has been the main cause of concern. The pseudo-meeting for money transaction, electoral run controversies, and pornographic scandals are results of this misuse.

Deepfakes possess a dual-use due to its accessible and unregulated nature. While the positive uses promote game-changing technological and societal benefits, deepfakes also enable the use of all the discussed malicious activities. Thus, the complexity behind the scenes and in the final product necessitates a nuanced method in regulation, considering intent, context, and impact.

3.3 Deepfake Detection: the Challenge & Methods

Deepfakes became increasingly sophisticated up to the point where a brief human perception alone may be insufficient for detection. Plus, due to the

nature of GANs with the detective, automated systems also struggle detecting manipulations. Even though it is still possible to tell the real from the fake with special attention to minor inconsistencies, the tool has become too convincing. A state-of-the-art example can be the Donald Trump deepfake video from section **1.3**, where one can be halfway into watching and reacting to it with the truth unbeknownst to them. Thankfully, there are established key details to consider while analyzing such deepfake media.

The first detail to focus on is the facial features, starting by the skin texture. Real human skin displays complex texture and variation that is yet to be perfectly captured by AI. Deepfaked faces often have a video-game-like mesh & texture. Unless perfected, the deepfake activity becomes evident. Hence, AI detecting systems can analyze inconsistencies & smoothness of the skin for deepfake detection. Next, the blinking patterns might also give the deepfake away. The algorithms might not correctly mimic natural subconscious human blinking behavior and eye movements, resulting in an eerie eye contact with the deepfaked subject. Moreover, the deepfaked face may show irregularities in head movements and facial orientation.

After the face, the observer can look for inconsistency in the subject's voice tone patterns. Although it is going through dramatic improvements, AI-generated speech may still fall short in perfectly replicating the natural shifts or stutters in tone and pitch. The next tip is interesting: background noise. A deepfaked video could be lacking subtle white noises that would mainly be present in genuine footage. Thus, also looking for ambient sounds and noises can deceive signs of manipulation activity.

Parallel to deepfake creation, deepfake detection is also continuously advancing. Increased sophistication in detection techniques are topics of research today. Funnily enough, AI is leveraging to combat deepfake media! Machine learning models are trusted to point out manipulation signs within more than meets the eye. Such models are good at identifying anomalies and specific patterns in vast datasets, highlighting the presence of deepfake usage.

The participation of Artificial Intelligence not only in the creation, but also in the goal for mitigation of deepfakes only concerns and motivates The EU AI Act more to become more robust and expansive in being a sufficient answer to the rise of deepfakes. Aside from deepfake creation assessment, automated deepfake detection tools should also be under consideration of The EU commission for further implementations of the AI Act. Further from the request of manual disclosure during the creation process, a suggested threshold for training data quality for detection algorithms will be much beneficial in imminent mitigation of misused deepfakes.

4. Conclusion

We've explored in detail the multifaceted controversies surrounding deepfakes, by considering the applications, regulatory challenges, and technological process throughout this thesis. While deepfakes are currently considered as a breakthrough development in technology with significant impact to society, it came with its benefits and doubts. Apart from appreciating its accessibility aid capabilities, the impersonation fraud side of it cannot be ignored.

There are significant ethical consequences of deepfake use that directly concern consensual, privacy, and misinformation risks. The advocacy of responsible use and development of the tool is needed to mitigate the potential of sinister events, ensuring that the benefits are actualized in an ethical manner. The legal frameworks concerned with this should be completed by strong standards and guidelines, promoting liability, transparency, and respect for human rights.

The EU AI Act pioneers a crucial step to regulate Artificial Intelligence and its risk potential; however, the current proposed framework has still some room for improvement in addressing the complexity of the deepfake technology satisfactorily. These improvements are needed in areas such as "high-risk" systems classification and ambiguous enforcement mechanisms to capture malicious AI content. An explicit inclusion of deepfakes under the "high-risk" title is necessary to ensure the tool is subject to thorough oversight and stricter requirements. Moreover, clear regulations that directly target the imminent removal of malicious deepfake content and liability of developers are necessary to avoid legal loopholes. Furthermore, the AI Act falls insufficient in keeping up with the ever-evolving nature

of AI and deepfakes, thus introducing a challenge tackling emerging threats. Following this analysis of the AI Act, the necessity of a revision was validated.

Additionally, since deepfakes pose a global risk, international cooperation in this deed for regulation will dramatically impact the effectiveness of the message. Countries must cooperate to harmonize their legal frameworks regarding AI and develop global standards and strategies. Synthetic media must be clearly labeled and adhere to consent and sensitivity. Plus, transparency measures to hold mispurposed developers accountable are unavoidable.

The future of Artificial Intelligence and deepfakes are still shaping as of today with swift advancements in technology and research. Therefore, ensuring the continuity and expansion of research & development in the detection of deepfakes is essential. Fortunately, the detection tools are becoming sophisticated enough to identify even bits of manipulation. Maintaining the parallelity in development of both the creation and detection of deepfakes will help keep pace to mitigate any misuse. New challenges will surely be introduced. Future technology might make it impossible to naturally identify deepfaked content, hence urging constant updating of regulations and detection methods. Therefore, scalable policies and innovative technological solutions will ensure for humanity to remain proactive and vigilant towards these challenges.

Overall, the multifaceted controversies surrounding deepfakes are best met by multifaceted approaches with ethical guidelines, robust legal frameworks, and continuous technological innovation. The benefits of the deepfake technology will be best harnessed by fostering global collaboration, promotion of responsibility in AI, and investment in mitigation strategies. Constant persuasion of adaptation and

vigilance will be key in ensuring ethical responsibility in the use of deepfakes in the future.

5. References

Barthle, C. (2023, April 27). *Humans + Machines: A Look Behind the Playlists Powered by Spotify's Algotorial Technology*. Spotify Engineering. Retrieved April 18, 2024, from https://engineering.atspotify.com/2023/04/humans-machines-a-look-behind-spotifys-algotorial-playlists/

Chen, H., & Magramo, K. (2024, February 4). *Finance worker pays out $25 million after video call with deepfake 'chief financial officer'*. CNN. Retrieved May 28, 2024, from https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

Gomes, B. (2024, May 14). *How Google's LearnLM generative AI models support teachers and learners*. The Keyword. Retrieved May 25, 2024, from https://blog.google/outreach-initiatives/education/google-learnlm-gemini-generative-ai/

Goujard, C. (2024, February 6). *Taylor Swift deepfakes nudge EU to get real about AI*. POLITICO.eu. Retrieved June 10, 2024, from https://www.politico.eu/article/europe-eye-fix-taylor-swift-nude-deepfake/

Jain, S. (2024, March 11). *Generative Adversarial Network (GAN)*. GeeksforGeeks. Retrieved June 9, 2024, from https://www.geeksforgeeks.org/generative-adversarial-network-gan/?ref=lbp#architecture-of-gans

Lana Del Rey - Skyfall (Adele) [AI cover]. (2023, May 10). YouTube. Retrieved April 20, 2024, from https://www.youtube.com/watch?v=t7jV5sSLWPc

Li, Y., Yang, X., Qi, H., & Lyu, S. (n.d.). *yuezunli/celeb-deepfakeforensics: Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics*. GitHub. Retrieved June 10, 2024, from https://github.com/yuezunli/celeb-deepfakeforensics

Li, Y., Yang, X., Qi, H., & Lyu, S. (2019, September 27). *[1909.12962] Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics*. arXiv. Retrieved June 10, 2024, from https://arxiv.org/abs/1909.12962

Murphy, S. (2024, May 17). *Many high schools are curbing the use of AI. These schools are leaning in*. CNN. Retrieved May 27, 2024, from https://edition.cnn.com/2024/05/17/tech/ai-high-school/index.html

Nobody Said This. (2024, May 16). Instagram. Retrieved June 1, 2024, from https://www.instagram.com/p/C6SGMGvOP9W/

Sebastian, M., & Jain, A. (2024, May 15). *AI and deepfakes blur reality in India elections*. BBC. Retrieved May 30, 2024, from https://www.bbc.com/news/world-asia-india-68918330

Shedding light on AI bias with real world examples. (2023, October 16). IBM. Retrieved June 4, 2024, from https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/

Spotify Premium Users Can Now Turn Any Idea Into a Personalized Playlist With AI Playlist in Beta — Spotify. (2024, April 7). Spotify Newsroom. Retrieved April 18, 2024, from https://newsroom.spotify.com/2024-04-07/spotify-premium-users-can-now-turn-any-idea-into-a-personalized-playlist-with-ai-playlist-in-beta/

Wolf, Z. B. (2024, January 24). *The deepfake era of US politics is upon us*. CNN. Retrieved May 29, 2024, from

https://edition.cnn.com/2024/01/24/politics/deepfake-politician-biden-what-matters/index.html