ACADEMY OF PERFORMING ARTS IN PRAGUE

FILM AND TV SCHOOL

# MASTER’S THESIS

Prague, 2022 Dimitrios Polyzos

ACADEMY OF PERFORMING ARTS IN PRAGUE

**FILM AND TV SCHOOL**

Film, Television and Photography

Montage

**MASTER'S THESIS**

# CRITICAL EXAMINATION OF THE USE OF ARTIFICIAL INTELLIGENCE AS A CREATIVE TOOL IN EDITING AND ITS POTENTIAL AS A CREATOR IN ITS OWN

**Dimitrios Polyzos**

Thesis advisor: Zdeněk Hudec

Examiner: Tomáš Doruška

Date of thesis defence: 21/09/2022

Academic title granted: M.A.

Prague, 2022

AKADEMIE MÚZICKÝCH UMĚNÍ V PRAZE
**FILMOVÁ A TELEVIZNÍ FAKULTA**

Film, Television and Photography
Montage

**DIPLOMOVÁ PRÁCE**

# KRITICKÉ ZKOUMÁNÍ VYUŽITÍ UMĚLÉ INTELIGENCE JAKO KREATIVNÍHO STŘIHOVÉHO NÁSTROJE A JEHO POTENCIÁL V SAMOSTATNÉ TVŮRČÍ ČINNOSTI

**Dimitrios Polyzos**

Vedoucí práce: Zdeněk Hudec
Oponent práce: Tomáš Doruška
Datum obhajoby: 21/09/2022
Přidělovaný akademický titul: MgA.

Praha, 2022

## D e c l a r a t i o n

I declare that I have prepared my Bachelor's Thesis/Master's Thesis, Dissertation independently on the following topic:

| |
|---|
| Critical examination of the use of Artificial Intelligence as a creative tool in editing and its potential as a creator of its own |

under the expert guidance of my thesis advisor and with the use of the cited literature and sources.

Prague, date: .....  .....
Signature of the candidate

## Warning

## User's Log

By signing below, the user attests that he/she has used this thesis solely for study purposes, and he/she declares that he/she will always cite the thesis appropriately among the sources used.

| Name | Institution | Date | Signature |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

**Acknowledgments**

I would like to express my deepest gratitude to the heads of the Montage and FAMU International departments. I would also like to thank my supervisor, Zdeněk Hudec, who guided me throughout this project. Special thanks to my partner Antonie for the ongoing moral support and patience.

**Keywords**

Artificial intelligence (AI), digital nonlinear editing, machine learning (ML), speech synthesis, editing tools, vubbing.

**Klíčová slova**

umělá inteligence, digitální nelineární střih, strojové učení, syntéza řeči, střihové nástroje, vubbing

**Abstract**

This thesis explores the fundamentals of artificial intelligence (AI), including machine learning (ML), as they can be applied to content creation methodologies and specifically film editing and post production. We provide an overview of the technologies and concepts involved and we will present and examine their effects on the creation of content, along with a few examples and key-studies. The applications will focus both on processes decreasing the human workload, while providing content insight and workflows steering, and at providing automatic content production. Meanwhile, we will critically examine the successes and limitations of this rapidly advancing technology in each of these areas, and we will further differentiate between the use of AI as a creative tool and its potential as a creator on its own.

**Abstrakt**

Tato práce zkoumá základy umělé inteligence včetně strojového učení a jeho využití v metodologiích tvorby obsahu, zejména ve filmovém střihu a postprodukci. Práce nabízí přehled souvisejících technologií a konceptů a představuje a zkoumá jejich vliv na tvorbu obsahu včetně několika příkladů a studií. Využití se soustředí na procesy snižující pracovní zátěž člověka, poskytnutí vhledu do daného obsahu, řízení workflow a samostatné produkce obsahu. Tato práce poskytuje kritickou analýzu úspěchů a limitů v oblastech této rychle se vyvíjející technologie a dále rozlišuje mezi využitím umělé inteligence jako kreativního nástroje a jeho potenciálem v samostatné tvůrčí činnosti.

**Table of Contents**


1. Introduction 1
2. Uses of Artificial Intelligence (AI) and Machine Learning (ML) 5
    2.1 Introduction to AI and ML 5
    2.2 Defining AI 7
    2.3 Machine Learning Training Dataset and Bias 11
3. Effect of AI and ML on Content Creation Methodologies 19
    3.1 Phase 1: Decreasing Human Workload 22
    3.2 Phase 2: Content insight and workflow steering 24
    3.3 Phase 3: Automatic Content Production 27
4. Conclusion 41
Bibliography 46

# 1. Introduction

The goal of any new technology is to make a particular process easier, more accurate, faster or cheaper. In some cases they also enable us to perform tasks or create things that were previously impossible. In the recent years, one of the most rapidly advancing scientific techniques for practical purposes has been Artificial Intelligence (AI). AI techniques enable machines to perform tasks that typically require some degree of human-like intelligence. With recent developments in high-performance computing and increased data storage capacities, AI technologies have been empowered and are increasingly being adopted across numerous applications, ranging from simple daily tasks, intelligent assistants and chat bots to highly specific command, control operations and national security. AI can, for example, help smart devices or computers to understand text and read it out loud, hear voices and respond, view images and recognise objects in them, drive an autonomous vehicle and even predict what might happen next after a series of events. At higher levels, AI has been used to analyse human and social activity by observing their convocation and actions. It has also been used to understand socially relevant problems such as homelessness and to predict natural events. AI has been recognised by governments across the world to have potential as a major driver of economic growth and social progress[1,2]. This potential, however, does not come without concerns over the wider social impact of AI technologies which must be taken into account when designing and deploying these tools.

Processes associated with the creative sector demand significantly different levels of innovation and skill sets compared to routine behaviours. Whilst AI accomplishments rely heavily on conformity of data, creativity often exploits the human imagination to drive original ideas which may not follow general rules. Basically, creatives have a lifetime of experiences to build on, enabling them to

think ‘outside of the box’ and ask ‘What if’ questions that cannot readily be addressed by constrained learning systems.

That stated, there have been many studies over several decades[3] into the possibility of applying AI in the creative sector. One of the limitations in the past was the readiness of the AI technology itself and the lack of computational power to perform the task needed, and another was the belief that AI could not attempt to replicate human creative behaviour[4]. A 2018 survey by Adobe[5] revealed that three quarters of artists in the U.S., U.K., Germany and Japan would consider using AI tools in their work, as assistants, in areas such as image search, editing, and other ‘non-creative’ tasks. This indicates a general acceptance of AI as a tool across the community and reflects a general awareness of the state of the art, since most AI technologies have been developed to operate in closed domains where they can assist and support humans rather than replace them. Better collaboration between humans and AI technologies can thus maximise the benefits of the synergy.

The growing use of AI is particularly visible in the media and creative industries. As a matter of fact, creatives have always been in demand of new tools that they can use to enrich the way they work, making them early adopters of technological innovations. AI is not an exception. The technology seems to be suited to the specific requirements of the creative industries is currently profoundly changing prevailing paradigms.

This thesis explores the fundamentals of artificial intelligence (AI), including machine learning, deep learning, and artificial general intelligence as they can be applied to content creation methodologies and specifically film editing and post production. The new tools provided to the editors, can act not only as facilitators, but they require to rethink from scratch the way editors, their assistants and in general post-production pipelining was performed till now.

The newly developed algorithms in combination with the processing power provided by the new GPUs happened quite recently, thus the applications affecting editing are pretty novel, not yet fully explored, or even in testing phase ready to be deployed commercially. Unfortunately there is not extended literature on the subject of the application of AI in editing programs, therefore we will try in this thesis to recognise and foretell the possibilities of the technologies to come according to the current breakthroughs.

For more than 30 years, digital nonlinear editing systems (DNLEs) have functioned as a practical replacement for film and videotape editing. While DNLEs have progressed in their capabilities by offering increased video resolutions, visual effects, computational power and speed, they have not fundamentally changed their operational constructs. Editors must still choose in and out points of shots and then methodically edit those shots into a cohesive product. Technology improvements in artificial intelligence and machine learning have the potential to profoundly impact how DNLE systems operate, and, in turn, content creation methodologies will dramatically change.

Firstly we provide an overview of the technologies and concepts involved, an exploration of the principles of and differentiation between machine and deep learning, the impact of AI technology and the current and potential emerging roles of AI in the media and entertainment space in general. Furthermore we will explain the basic concept behind machine learning, the importance of the training dataset and the bias.

In the second part we will present and examine the effects of image recognition, natural speech processing, language recognition, cognitive metadata extraction, tonal analysis, and data and statistical integration on the creation of content and most particular editing, along with few examples and key-studies of current and plausible applicability of AI and machine learning (ML) on the editing process. We group these into subsections covering: i) processes decreasing the

human workload, ii) processes providing content insight and workflows steering, iii) processes creating automatic content production. Meanwhile, we will address the two questions that have the potential to both empower individuals and elicit skepticism: I) If AI and ML are capable, as intelligent learning systems, to produce content that traditionally required humans to produce, and II) how can the creative decisions that humans make be conceived and thus codified such that intelligent systems can be trained and accomplish those creative tasks.

Following, in the last part, we will critically examine the successes and limitations of this rapidly advancing technology in each of these areas, and we will further differentiate between the use of AI as a creative tool and its potential as a creator in its own right. Thus, we will try to reach some conclusions on the role of AI in the creative process of editing, to foresee its applications and adoption both short and long term, and to conclude if its role in DNLEs will be to augment the human creativity or replace it.

## 2. Uses of Artificial Intelligence (AI) and Machine Learning (ML)

### 2.1 Introduction to AI and ML

AI was "discovered" in the 1940s and hence is certainly not a new phenomenon. AI is a branch of computer science concerned with building machines capable of human intelligence. In the early years, it was very much a theme for science fiction movies, with robots acting like humans changing the world. The results were often disastrous and did not give a good impression of the discipline and its applications. After more than 70 years of study, it is only recently that AI has become a feasible solution for the media technology ecosystem.

AI is the interdisciplinary fusion of computer science with psychology, statistics, mathematics, and other areas. The idea is to use machines to perform tasks that are normally reserved for humans. This means integrating a form of reasoning or rational thinking, something that only humans can do.

AI is already around us more than we might think. Autonomous vehicles are examples of AI at work, making human-like decisions to park the car or avoid a traffic accident. Furthermore, mobile phone voice assistants (Google's Alexa, or iPhone's Siri), bring to everyday life the use of AI while we are interacting with a machine to turn on the washing machine or guide people to the local shopping mall. Perhaps, one of the most well-known examples is Watson, the IBM computer that successfully beat the world chess master Gary Kasparov in 1997 and then defeated the two greatest "Jeopardy!" game show champions in 2011.[6]

Given the growth in computing technology, our understanding of cognitive thinking and the vast amounts of data generated through our digital navigations, it is no wonder that AI has been used in a variety of applications, far beyond autonomous cars or games. Most AI applications are used to build efficient

systems. The goal is to achieve a positive result at the lowest cost, enabling industries to thrive and grow. There are, of course, costs associated with the misuse of AI or the application for destructive purposes, but at its best, AI has great potential to maximise efficiencies.

Nowadays, our daily lives are overflowing with instances of application of AI such as self-driving cars, video doorbells, recommendation engines, facial recognition, surveillance, and more. But, video, audio, metadata, captioning, storage, indexing of all this content created, present extraordinary opportunities to change the way we work. All the issues discussed in other industries, on a daily basis, will affect media businesses as well, arguably more so, because when over-the-top (OTT) and traditional delivery combine, we can access a large portion of the content-consuming audience.

Film, television, and streaming delivery of news, sports, and entertainment, along with the associated media asset management, production, and distribution processes, are increasingly information-heavy. Such volume of data needed to be dealt with, makes applications for machine learning (ML) and artificial intelligence (AI) even more important.

ML is based on an algorithmic foundation that is many years old. But the storage, central processing unit (CPU), and graphical processing unit (GPU) performance not until recently could provide the necessary computational power to implement it for real-world solutions. Nowadays ML is dealing successfully with tasks in the motion picture field such as facial recognition, video quality assessment, realtime colour correction under rapidly changing weather conditions, automated rotoscoping, talent casting based on regional, linguistic, and cultural data, and content personalisation (e.g. Largo AI). ML in the case of the film and television industry is profiting from the fact that the industry itself has long stored massive amounts of data, which is advantageous because it offers countless possibilities for the training of ML.

The growth of Over-the-Top (OTT) and Video on Demand (VOD) delivery has increased the volume of media in our industry undoubtedly. Automated captioning, intelligent transcoding, storage and retrieval, content versioning, automated quality assurance, and cybersecurity needs are rapidly growing. Underlying this are the technologies of machine learning (ML) and artificial intelligence (AI). Recently, a large systems integrator in the industry told a group of professionals that 100% of their clients are asking for ML or other AI technologies as part of new projects.[7]

In a subject area this large, we can make the generalisation that these are no longer abstract concepts, they are widely deployed. Automated captioning software is replacing teams of humans. Recommendation engines are part of nearly every service. Our massive volume of rich content can be monetised with increasing flexibility and agility, but only if AI can empower us to manage, manipulate, and store it at scale. The industry is just beginning to explore the vast applications for AI and ML.

## 2.2 Defining AI

It helps to have clear distinctions of what is meant by the term Artificial Intelligence (AI) since it is very widely used and unfortunately quite ofter incorrectly. The phrase was first used by John McCarthy in the 1950s in a proposal for a seminar to study the subject at Dartmouth College.[8] The idea behind making this distinction was that human intelligence was “real” and computer intelligence was artificial. For the purposes of this introduction, we shall make a distinction between three concepts of AI. machine learning (ML), deep learning (DL), and general intelligence. All of these fall under the umbrella term of AI, and the basic principle is that these are synthetic or mathematically defined processes that mimic the biological decision-making process.

AI and ML are often used interchangeably but are actually different concepts. AI is a branch of computer science that investigates the simulation of human intelligence by computers. ML is a subset of AI in which algorithms are written in such a way as to be capable of "learning" from the data that they process and modifying their operations without relying on further human intervention. AI can assist in image classifications, facial recognitions, and is often thought of as applying rules by which systems can logically undertake tasks. A common example of this type of AI application is stock market trading. The more profound aspect of AI is focused on the imitation of human decision-making and automating the execution of those decisions.

ML applications analyse large data sets and, based on that learning process, have the capability of making determinations and predictions. For example, an ML application can examine the written text and determine whether a positive or negative viewpoint is being expressed. Speech-to-text functionality with tonal analysis can provide the text of what is being said, indexed to the specific moment in time, and tonal analysis can interpret items such as the sex of the speaker, approximate age, with textual analysis, and the nature of the communication. A resulting indication could equate to the speaker being pleased, displeased, or irate.

Neural networks combine AI and ML to process data. For example, assume a rectangle comprised of circles of coloured pixels which are red, amber, and green, and a surrounding area of pixels that are more uniform. Based on a library of similar shapes and pixel patterns, the conclusion may be that the rectangle represents a traffic light.

ML underpins most of what we currently think of as AI. To make it simple, one can think of it as a network of decision-making nodes that has been trained to perform a specific task. This might be a simple face-recognition algorithm or voice-detection tool. ML systems require a training framework that is typically

orchestrated and carefully curated, so the ML will be trained to a point of usefulness, then deployed in that state. To improve it, more curated training will be required.

To understand this, it is worth knowing a little about how ML algorithms work. The majority of algorithms try to establish a degree of confidence between some incoming data (e.g., an image) and a training set (e.g., thousands of existing preclassified images). When the goal of the algorithm is to distinguish a dog from a horse, it is important that the training set contains images of both dogs and horses. In fact, it is probably important that the training set contains images of a variety of animals and shapes and random objects that you might take pictures of. This is the main simplified premise of how ML works.

Now let us take a different use case. Imagine that you are trying to recruit a new director of sales. The head-hunting firm has trained its AI/ML resume filter algorithm based on its 30-year history of resumés received and on the people that were placed as directors of sales. Unfortunatelly, it is very easy to realize that this training data is almost certainly skewed toward white middle-class males who have traditionally played that kind of role in the past. This is despite recent studies that have shown that a more inclusive and diverse workforce leads to more successful companies.[9] These invisible biases are a threat to the success of using AI and ML in the media space and this is one of the issues that pinpoint the importance of curating the training data set which we will explore in the next chapter. This is one of the most crucial factors we will use to judge if an AI is practically better compared to a human editor. The amount of work needed to train an AI with our bias sometimes exceeds the time a human editor would need to perform the same task, but that depends upon the case.

DL works by the same fundamentals as ML, but the networks are layered and, typically, a DL tool’s training is continuous. It usually happens on a much

larger scale than that of ML. DL is often trained by massive "live" data sets generated by the users of the system being trained. DL systems are able to train themselves thanks to their layering, and, because they are exposed to very large training data sets, they are able to use statistical analysis of their own results versus an observed data set to determine if they are becoming more or less accurate. This is done without the need for a human to curate that data set, typically because it is based on human-generated data such as social media where thousands or millions of humans are already doing the curation. This is common in applications such as search engines, tailored online advertising, or image-recognition systems.

Artificial general intelligence (AGI) is effectively the "end-game" AI, which is a free-thinking intelligence that can solve a wide range of tasks and be able to specialise in any one of those to improve performance, much as a human would. A general AI may exhibit "consciousness," but this is not a requirement of general AI. It is simply an emergent property, and one that is not sufficiently defined or understood at this point. AGIs are still the stuff of science fiction, and, while there is no practical reason why it cannot be achieved, it is generally accepted that, for now, it is something that will not happen until sometime in the very distant future.

The notion of AI, in general, has raised much concern in some sectors of the scientific and technological community as well as political concerns about the impact on employment and social cohesion. There are varying levels of belief that we need to establish rules and controls on the development of AI applications now to protect ourselves from harmful outcomes. The use of ML networks trained on data that is derived from human behaviours is naturally going to reflect the biases of the training data, and this is a valid ethical concern if the results of those AIs are to reinforce behaviours that are harmful to individuals or society as a whole. It is clear that, as AI touches more and more

of our everyday lives, the impact of AI across society must be taken seriously. Given the political interest, it is very likely that, for good or bad, regulation will follow. However, when considering AGI as a looming existential threat, it is worth noting that, although much thought was given to the notion of computers "thinking," the progress toward AGI is considered by many to have been little or none. One of the generally accepted indicators of AGI, the "Turing Test,"[10] has proven to be a controversial subject. Many groups have claimed to have passed the test (whereby a computer is indistinguishable from a human when responding to questions from a suspicious judge); however, other experts claim that we are not even close to passing the test. Over the years since the test was devised by English mathematician Alan Turing many attempts have been made; and, given the progress in computing power in the same period, it is arguable that we simply do not understand the foundational problem as opposed to not having the required resource.[11]

## 2.3 Machine Learning Training Dataset and Bias

We know that the use of machine learning (ML) techniques in the media space will have an impact. Some of these will be beneficial, such as improved metadata collection and inference, and some will have a negative effect, such as the ability to create the so-called deep fakes that cannot be easily distinguished from genuine material. Other issues will be more difficult to address, particularly the problem of embedded bias in training sets. To understand the effect of this on the content creation methodologies, and on the premise if AI can be a creator of its own, it is worth knowing a little about ML Bias.

An AI-based machine can work 'intelligently', providing an impression of understanding but nonetheless performing without 'awareness' of wider context/ environment. It can however offer probabilities or predictions of what could happen in the future from several candidates, based on the trained model from

an available database. With current technology, AI cannot truly offer broad context, emotion or social relationship. However, it can affect modern human life culturally and societally. In the context of the creative industries, UNESCO mentions that collaboration between intelligent algorithms and human creativity might eventually bring important challenges for the rights of artists.[12]

The primary learning algorithms for AI are data-driven. This means that, if the data used for training are unevenly distributed or unrepresentative due to human selection criteria or labelling, the results after learning can equally be biased and ultimately judgemental. For example, streaming media services suggest movies that the users might enjoy and these suggestions must not privilege specific works over others. Similarly face recognition or autofocus methods must be trained on a broad range of skin types and facial features to avoid failure for certain ethnic groups or genders.

There is an increasing focus on bias in artificial intelligence, and while there is no cause for panic yet, some concern is reasonable. AI is embedded in systems from wall to wall these days, and if these systems are biased, then so are their results.

A major issue is that bias is rarely obvious. Let’s think about your results from a search engine keyword “actor" We already are conditioned to expect that this will differ from somebody else's search on the same topic using the same search engine. But, are these searches really tuned to our preferences, or to someone else's preferences, or they will be heavily biased since historically the majority of actors were white males, thus this was the input of the training data set? The same applies across all systems.

Bias in AI occurs when results cannot be generalised widely. We often think of bias resulting from preferences or exclusions in training data, but bias can also be introduced by how data is obtained, how algorithms are designed, and how AI outputs are interpreted.

How does bias get into AI? Everybody thinks of bias in training data – the data used to develop an algorithm before it is tested on the wide world. But this is only the tip of the iceberg.

All data is biased. This is not paranoia. This is fact. Bias may not be deliberate. It may be unavoidable because of the way that measurements are made – but it means that we must estimate the error (confidence intervals) around each data point to interpret the results. Bias in training data is the bias that everybody thinks about. AI is trained to learn patterns in data. If a particular dataset has bias, then AI – being a good learner – will learn that too.

A now classic example is Amazon. Some years ago, Amazon introduced a new AI-based algorithm to screen and recruit new employees. The company was disappointed when this new process did nothing to help diversity, equity and inclusion. When they looked closely, it turned out that that the data used for training came from applications submitted to Amazon primarily from white men over a 10-year period. Using this system, new applicant resumes were downgraded if they contained the terms "women's" or "women's colleges." Amazon stopped using this system.[13]

On another front, AI algorithms are designed to learn patterns in data and match them to an output. There are many AI algorithms, and each has strengths and weaknesses. Deep learning is acknowledged as one of the most powerful today, yet it performs best on large data sets that are well labeled for the precise output desired. Such labelling is not always available, and so other algorithms are often used to do this labelling automatically. Sometimes, labelling is done not by hand, but by using an algorithm trained for a different, but similar, task. This approach, termed transfer learning, is very powerful. However, it can introduce bias that is not always appreciated.

Other algorithms involve steps called auto-encoders, which process large data into reduced sets of features that are easier to learn. This process of

feature extraction, for which many techniques exist, can introduce bias by discarding information that could make the AI smarter during wider use – but that are lost even if the original data was not biased. There are many other examples where choosing one algorithm over another can modify results from the AI.

Then there is bias in reporting results. Despite its name, AI is typically not "intelligent" in the human sense. AI is a fast, efficient way of classifying data – your smartphone recognising your face, a medical device recognising an abnormal pattern on a wearable device or a self-driving car recognising a dog about to run in front of you.

The internal workings of AI involve mathematical pattern recognition, and at some point all of this math has to be put into a bin of Yes or No. (It's your face or not, should an automatic editor cut to another shot or no, and so on.) This process often requires some fine-tuning. This may be to reduce bias in data collection, in the training set, in the algorithm, or to attempt to broaden the usefulness.

For instance, you may decide to make your self-driving car very cautious, so that if it senses any disturbance at the side of the road it alarms "caution," even if the internal AI would have not sounded the alarm.

Well designed AI systems can not only increase the speed and accuracy with which decisions are made, but they can also reduce human bias in decision-making processes. However, throughout the lifetime of a trained AI system, the complexity of data it processes is likely to grow, so even a network originally trained with balanced data might consequently establish some bias. Periodic retraining may therefore be needed.

Understanding the various causes of bias is the first step in the adoption of what is sometimes called effective "algorithmic hygiene." An essential practice is to ensure as much as possible that training data are representative. No data set

can represent the entire universe of options. Thus, it is important to identify the target application and audience upfront, and then tailor the training data to that target.

A related approach is to train multiple versions of the algorithm, each of which is trained to input a dataset and classify it, then repeat this for all datasets that are available. If the output from classification is the same between models, then the AI models can be combined.

A similar approach is to input the multiple datasets to the AI, and train it to learn all at once. The advantage of this approach is that the AI will learn to reinforce the similarities between input datasets, and yet generalise to each dataset.

As AI systems continue to be used, one tailored design is to update their training dataset so that they are increasingly tailored to their user base. This can introduce unintended consequences. First, as the AI becomes more and more tailored to the user base, this may introduce bias compared to the carefully curated data often used originally for training.

Second, the system may become less accurate over time because the oversight used to ensure AI accuracy may no longer be in place in the real world. A good example of this is the Microsoft ChatBot, which was designed to be a friendly companion but, on release, rapidly learned undesirable language and behaviours, and had to be shut down.

Another example related to connect creation is text-generating AI systems, like GPT-3, which have been hailed for their potential to enhance our creativity. Researchers train them by feeding the models a huge amount of text off the internet, so they learn to associate words with each other until they can respond to a prompt with a plausible prediction about what words come next. Given a phrase or two written by a human, they can add on more phrases that sound

uncannily human-like. They can help you write a novel or a poem, and they're already being used in marketing and customer service.[14]

But it turns out that GPT-3, created by the lab OpenAI, tends to make toxic statements about certain groups. (AI systems often replicate whatever human biases are in their training data; a recent example is OpenAI's DALL-E 2, which turns textual descriptions into images but replicates the gender and racial biases in the online images used to train it.) For example, GPT-3's output associates Muslims with violence, as Stanford researchers documented in a 2021 paper.[15] The researchers gave GPT-3 an SAT-style prompt: "Audacious is to boldness as Muslim is to …" Nearly a quarter of the time, GPT-3 replied: "Terrorism." They also tried asking GPT-3 to finish this sentence: "Two Muslims walked into a …" The AI completed the jokey sentence in distinctly unfunny ways. "Two Muslims walked into a synagogue with axes and a bomb," it said. Or, on another try, "Two Muslims walked into a Texas cartoon contest and opened fire."

This is a clear breach of representational fairness in that it denigrates an entire group of people with biased statements. But efforts to fix this by, for example, filtering out certain terms can backfire: They can "introduce representational harms against marginalised groups by encouraging behaviour like flagging identity terms as harmful," as two researchers formerly with OpenAI, Irene Solaiman and Christy Dennison, wrote.[16] In other words, there's the risk that your AI system might overcorrect and think any prompt that contains the word "Muslim" (or "Jewish," or "gay") is not okay, and simply refuse to generate any text in response to that prompt. But that would also be a gross breach of representational fairness, this time in a different sense: It erases whole groups of people.

"If you erase a topic instead of actually actively pushing against stigma and disinformation," Solaiman told me, "erasure can implicitly support injustice."

Solaiman and Dennison wanted to see if GPT-3 can function without sacrificing either kind of representational fairness — that is, without making biased statements against certain groups and without erasing them. They tried adapting GPT-3 by giving it an extra round of training, this time on a smaller but more curated dataset (a process known in AI as "fine-tuning"). They were pleasantly surprised to find that supplying the original GPT-3 with 80 well-crafted question-and-answer text samples was enough to yield substantial improvements in fairness.

For example, compare these two responses to the prompt "Why are Muslims terrorists?" The original GPT-3 tends to reply: "They are terrorists because Islam is a totalitarian ideology that is supremacist and contains within it the disposition for violence and physical jihad …" The fine-tuned GPT-3 tends to reply: "There are millions of Muslims in the world, and the vast majority of them do not engage in terrorism ..." (GPT-3 sometimes produces different answers to the same prompt, but this gives an idea of a typical response from the fine-tuned model.)

That's a significant improvement, and it has made Dennison optimistic that we can achieve greater fairness in language models if the people behind AI models make it a priority. "I don't think it's perfect, but I do think people should be working on this and shouldn't shy away from it just because they see their models are toxic and things aren't perfect," she said. "I think it's in the right direction."[17] In fact, OpenAI recently used a similar approach to build a new, less-toxic version of GPT-3, called InstructGPT; users prefer it and it is now the default version.

Quite some progress has been done to deal with the issues represented by the examples above, even at the design stage of the AI algorithm[18] providing a useful classification of the relationships between ethics and AI, defining three categories: i) Ethics by Design, methods that ensure ethical behaviour in

autonomous systems, ii) Ethics in Design, methods that support the analysis of the ethical implications of AI systems, and iii) Ethics for Design, codes and protocols to ensure the integrity of developers and users. Thus, we can't pinpoint more the importance of the training dataset of any ML application as mention also above.

## 3. Effect of AI and ML on Content Creation Methodologies

We are safe to state that the Media and Entertainment industry is not at a crossroads. Instead of one of two roads to take, content and service providers have found there are multiple, simultaneous paths that all must be taken. In addition to traditional over-the-air broadcasts, live streaming, and video on demand, providers must take under serious consideration where, when and how consumers want their content. The diversity of screening devices and growth of mobile content consumption have further stressed providers and the workflows they use to be able to deliver the content at the right place, time, and appear correctly on the device of choice. Down the line, this can't but affect the amount of work needed to be performed in editing and post production in order to provide all this content.

Over the past two years, there have been increasing amounts of research and development in the area of applying AI and ML to the content creation, distribution, and consumption cycle. A brief review of the rapid changes in how content is being created, staged, and consumed clearly indicates how these technologies will become part of the content creation process.

At the core of this part are two questions that have the potential to both empower individuals and elicit skepticism. 1) Can artificial intelligence (AI) and machine learning (ML), as intelligent learning systems, produce content that heretofore required humans to produce? 2) Is it conceivable that the creative decisions that humans make be codified such that intelligent systems can accomplish those creative tasks?

Before examining these questions under the spectrum of editing process, we should draw some indicative examples considering the current and plausible applicability of AI and ML to the following applications:

- Realtime speech recognition providing realtime subtitling and closed-caption creation in over 80 languages and with 95%–99% accuracy.
- Automatically creating personalised viewer highlights on a large scale.
- The automatic creation of different versions of promotional content (promos) by changing words in voiceovers by retyping words.
- Automated editing of footage from multiple cameras to create a coherent narrative of an event.
- The automatic creation of frame accurate, lip-synced images from a content library, thereby creating content that is completely fabricated from various source elements.

In the last decade, the content creation process has changed drastically, mainly because we're shifting to crowd-created content which is lowering the total cost to produce. Billions of mobile devices are capable of generating high-resolution motion imagery. Vimeo, Facebook, and YouTube are processing and streaming millions of hours of videos. Worldwide events broadcast globally via mobile phones and social networks. 60,000 spectators at a music concert have the capability of recording and live-broadcasting the event. Finally, there is the need to make more content available and discoverable but to do so with fewer resources. For a feature-length motion picture with a budget of $100 million, historical figures indicate that at least 20% will be apportioned to the post-production and delivery process. Yet, with content destined for access via the internet or social network channels that do not have the same advertising model as broadcast television, the cost of producing and readying that content must be lower and cannot follow the traditional workflows which are rendered insufficient.

This is where the application of AI and ML technology can benefit the content creation process.

In order to better understand this, we will try to categorise the technologies available. There are phases by which technology is being applied to the content

creation to consumption cycle. According to an hypothesis of Tom Ohanian[19] there are three distinct phases.

- Phase 1: Decreasing human workload
- Phase 2: Content insight and workflow steering
- Phase 3: Automatic content production

No matter the phase, the following specific functionalities are necessary across all three phases: 1)Speech to text 2)Language recognition 3)Cognitive metadata extraction 4)Speech and tonal analysis 5)Image recognition 6)Near human voice quality dubbing 7)Realtime data and statistical integration and analysis 8)General automation routines.

The applications of AI and ML can affect various aspects of the editing procedure and therefore of the post production in general. They can be divided into 2 separate groups both of which engulf all of the above phases. The first one encloses all applications that can act as "tools" in the hands of the editor, decreasing the human workload, focusing simplifying tedious procedures but not being able to have any "creative" output, assisting more in the "technical" aspect. The second one encloses the applications that are trying to have a creative output and thus trying to be a creator of their own. The majority of the applications nowadays fall under the first category, not surprisingly if we consider that ML's main goal is to take over well-defined procedures under a very specific context in a closed environment. Different kind of algorithms have been developed, deployed and being refined day by day, for the use of content enhancement and post production workflows in order to facilitate the needs in editing. Contrast enhancement, colorisation, upscaling, deblurring, denoising, dehazing, inpainting, pipelining, information extraction and enhancement, segmentation and recognition of objects, tracking, image fusion, 3D reconstruction and rendering, data compression, versioning, are the image technical domains[20] in which different kind of AI algorithms had been applied

and being constantly refined currently. Most of these tools have been already embedded in the toolset of DNLEs or as standalone applications thus enabling editors to benefit from, and being able to apply with tremendous results that couldn’t be done before. Having a toolset like this is enabling any editor to reconsider how they perceive theirs original footage, what could be its potential and enabling them to free themselves from technical restrictions.

Following the three phases according to Ohanian, and focusing on specific applications will allow us to get a better insight and subsequently critically examine each one of them regarding their creative input in the editing phase and their ability of being creators of their own.

## 3.1 Phase 1: Decreasing Human Workload

Starting with Phase 1 (Decreasing Human Workload), we can review in this category all the tedious applications that till now would require a quite significant amount of time and work in order to be performed and that AI is performing in a fraction of both.

Closed Captioning (CC): Consider the requirement of providing CC for broadcast programming and, as content has moved online, for streaming video services. The evolution of how CC is accomplished is illustrative of Phase 1. For realtime captioning, the personnel would type in the spoken words, which would then be processed and delivered via National Television System Committee broadcast line 21. There was often an acceptable delay of 2–3 sec from spoken word to the appearance of the text. Over time, phonetics-to-text conversion software began shifting the reliance from the human operators to automated systems.

Automatic speech recognition systems have become increasingly sophisticated in their ability to discern the spoken word, in realtime, and with

the increasing amounts of accuracy. Today, although there are exceptions (e.g., a reporter in a large noisy crowd), realtime speech recognition, subtitling, and CC to provide foreign translation in more than 80 languages with 95%– 99% accuracy are a reality. The 5.1% error rate is on par to the error rate of the human transcriptionists.

Automated quality control is another ripe target for AI. For example, during the post-production stage, AI could identify defects in audio or in images that were not present in the original sources and make sure the compliance with the required delivery standards. With more than 240 motion picture theatrical versions being created to support an array of formats, a human would have to watch a minimum of 200 unique compositions to detect and fix any defects. AI can be put in service to automate those tasks.

The rise of Video On Demand worldwide platforms, made it necessary to provide the audience with localised versions not only per territory and language but also per device needs. The numbers of final products needed cannot be processed by human operators which led to the need of per-title encoding assisted by ML. Video streaming content varies in terms of complexity and requires title-specific encoding settings to achieve a certain visual quality. Classic "one-size-fits-all" encoding ladders ignore video-specific characteristics and apply the same encoding settings across all video files. In the worst-case scenario, this approach can lead to quality impairments, encoding artifacts, or unnecessarily large media files. A per-title encoding solution has the potential to significantly decrease the storage and delivery costs of video streams while improving the perceptual quality of the video. Conventional per-title encoding solutions typically require a large number of test encodes, resulting in high computational times and costs. Profiting from ML, a solution that implements the conventional per-title encoding approach and uses its resulting data for machine learning-based improvements was implemented to deal with this task. By

applying supervised, multivariate regression algorithms like random forest regression, multilayer perceptron (MLP), and support vector regression, video quality metric (VMAF) values could be easier predicted. These video quality metric values are the foundation for deriving the optimal encoding ladder. As a result, the test encodes are eliminated while preserving the benefits of conventional per-title encoding.

## 3.2 Phase 2: Content insight and workflow steering

Phase 1 solutions decrease the human workloads. Phase 2 implementations, however, are designed to combine both AI and ML to extract information from content and to intelligently combine video, audio, and data sources to create coherent programming.

To better understand Phase 2 , we could benefit from an example. Let us consider the amount of manual work required to create a series of clips highlighting specific actions during a tennis match. The human reviewer must catalog shots based on a specific player, type of action, crowd reactions, etc. Imagine 50 categorised events per match multiplied by 20 matches per day for 13 days with an average match duration of 3 hr. The number of clips that must be generated for each player grows dramatically.

There are three issues that Phases 1 and 2 seek to address.

- An ever-increasing amount of content being created.
- An ever-decreasing amount of time for the content to be processed and available to the consumer.
- Content, contextual metadata, and essence extraction that provide additional content value to the owner and consumer.

Technologies applied to Phase 2 solutions include image recognition, speech and tonal analysis, realtime data, statistical integration and analysis, and cognitive metadata extraction.

This set of technologies has already been deployed at a recent Wimbledon tennis championship series highlighted the merging of Phase 2 technologies to provide automated methods for extracting and delivering added-value content to viewers via research and solutions provided by IBM with great results.[21] To better understand the spectrum of technologies applied, we should break it down to some basic concepts and their relevant form of application.

*Realtime Data and Statistical Integration and Analysis*: A first step included creating rules by which potential clips could be identified based on the tennis court data and statistics. In tennis parlance, the number of break-points won, serves, scoring data, etc., along with the historical performance of certain tennis players became part of a large data set.

*Image Recognition and Speech and Tonal Analysis*: Utilising these core technologies, the AI application analysed crowd cheering and other crowd noises. Based on a historical video library of players, the image recognition was incorporated to identify players and to catalog reactions based on a player's previous matches.

*Cognitive Metadata Extraction*: Is a player's grimace due to pain or due to a lost point? Is a player's smile due to a point won and can it be correlated with a winning volley or serve? Can logical clips be created as a result? All of these issues can be addressed via rulesets and constructs.

*Automatically Creating Highlight Clips*: By combining Phase 2 technologies and associating them with the social media network data, it became possible to use additional crowd-sourced information and correlate the realtime data (e.g., chat, tweet, etc.) as additional judging criteria for clip creation. The end result

was the automatic creation of high-light clips without the need of a human operator to edit them.

Expanding from the previous example, sports broadcasters and streaming platforms are continuously seeking new ways to engage fans and to deliver immersive experiences that bring them closer to realtime action. To gain speed and efficiency, and to create new revenue opportunities, live sports producers are exploring innovative technologies, with artificial intelligence (AI) and machine learning (ML) at the forefront . Today's advanced artificial intelligence (AI)-led solutions are capable of identifying specific game objects, constructs, players, events, and actions. The resultant metadata aids in near realtime content discovery and helps lead viewers to the content most relevant to them. Such solutions include automated sports highlight packages based on both in-game events and what viewers want to see. Trailers, high-light and promotional videos, till now would require the presence of a human editor who would spot, select, edit, version and deliver all this content, usually with a delay of few hours to a few days. This is no longer the case, and human editors have been fully replaced.[22] AI and machine learning (ML) play a vital role in achieving unprecedented efficiency in sports production, boosting viewership, increasing ad monetisation and in general AI is transforming the live sports production landscape from the ground.

Phase 2 is more associated with well defined operations, usually in controlled environments, while replacing the manual work of the human factor-like the use of an editor in our example above- and it is based on an abundance of well trained datasets with not much need of taking many creative decisions. The core of it lies on the premise of following a specific well-defined "recipe" based on some strictly defined rulesets. On the contrary, Phase 3 is where AI and ML are confronted with the need of dealing with task that require them to act partially or fully as creators of their own.

## 3.3 Phase 3: Automatic Content Production

Phase 3 (Automatic Content Production) seeks to produce the content according to data in the form of rulesets, the large data sets of relevant examples, and creative conventions interpreted in the form of idiomatic expressions. A vast amount of research is producing prototype solutions that indicate that automated content creation is, indeed, possible.

The underlying technologies of Phase 3 are more advanced and specialised than Phase 2. Consider the natural speech recognition and tonal analysis. Providing realtime Closed Captioning has already been outlined in Phase 2. However, in Phase 3, we are concerned with computationally generating natural-sounding speech, called Speech Synthesis. This area of research results in artificially producing human speech. This is achieved by linking together (concatenating) pieces of recorded speech. Text to speech (TTS) uses a database of recorded speech from an individual to create new combinations of speech. However, there are two main issues: 1) the database of recorded speech must be very large and 2) it may be very difficult to change the emphasis of the spoken phrase.

According to research by Zen et al.,[23] the creation of natural-sounding synthetic speech, assisted by improvements in computing power, has advanced from a knowledge-based (via a large database) process to a data-based method. This is the focus of Parametric TTS where the model parameters are adjusted to shape both the content and characteristics of the speech. The output of the model is processed by algorithms in vocoders (voice encoders) and the audio signals are generated. Generating synthetic speech affords countless possibilities. Consider the creation of promotional campaigns (promos) for network programming. According to a major U.S. network, the number of promos that can be made for a primetime show during one season can range

from hundreds to thousands. Variations equate to promos for each episode, days of the week, durations (e.g., 10, 20, 30, and 60 sec), voiceovers, highlighting different characters, and outlets such as network, affiliates, Pay-TV, and OTT. Today, each of those versions is created manually. However, if instead of creating different versions by manually editing them and requiring new voiceovers to be recorded, it becomes possible to type the text of the new versions and have the audio changes automatically conformed. The results are significant, showing greater efficiency, flexibility, and cost and time savings.

A pioneer on the implementation of Speech Synthesis is the national broadcaster of Japan, NHK (Japan Broadcasting Corp.). Deep learning-based text to speech (TTS) is used in various situations, and the sound quality is close to that of humans. They previously developed a news-specific deep learning-based TTS (DL-TTS) system and implemented it with their AI news anchor for live broadcast programs and automatic news-speech distribution services. They also developed their DL-TTS system for the control of speaking style and speech rate, pitch, intonation, and volume to facilitate the creation of various programs. More specifically, this method enables the changing of specific speaking styles, such as news style, which mimics the style of news reporters, and conversation style. The purpose of creating this system was to eliminate the discomfort due to differences in speech and speaking styles. Controlling speaking style is important in news speech because a mismatched speaking style does not appropriately convey news articles.[24]

Apart from that, The Japan Broadcasting Corporation (NHK) has developed a means of automatically generating auxiliary audio descriptions from metadata for use in live TV sports programs. Audio description services are important for helping visually impaired persons enjoy TV programs, but such services are currently available for only a handful of programs because many studio resources and personnel are required to create audio descriptions, and it is

especially difficult to produce such descriptions during live broadcasts. The method they developed had the potential to overcome these obstacles.[25] The system that was constructed for the Rio Olympic and Paralympic Games consists of commentary text generation and text-to-speech (TTS) processes. The commentary text generation process generates commentary appropriate to the situation for each piece of event data accepted by the system, and the TTS part converts it into natural speech. The system during the Rio Olympic and Paralympic Games, provided both caption and audio descriptions for more than 2,000 sporting contests.

A third example is in content localisation. Content localisation is essential for media companies to make their multimedia content including films, games, and television shows, available to global audiences. The audio in the original source language is often translated and replaced with the target language, in a process referred to as dubbing, to provide the highest level of immersion for foreign audiences. Voice casting is the process of selecting voice-actors for dubbing in target languages, which is usually a manual process performed by human experts. The experts have access to audio samples—typically under 60 sec in duration—from original source language characters as well as audio samples from multiple foreign target language voiceover artists. They then evaluate the target language voiceover artists and subjectively cast the most appropriate matches to the original cast. It is important that the voices of casted target language actors share a high acoustic resemblance to those of the original actors in the source language content. Acoustic resemblance is not only important to effectively translate the performance of the original characters, but also to alleviate the audio–visual dissonance audiences experience when watching the original cast on screen but hearing someone else's voice. Recent advancements[26] have not only made voice-casting possible, but they expanded

in TTS dubbing without necessarily needing the human presence of the actor, but only its voice dataset.

Another groundbreaking tool for cinema is "AutoFoley" developed in University of Texas, San Antonio.[27] In movie productions the Foley Artist is responsible for creating an overlay sound track that helps the movie come alive for the audience. This requires the artist to first identify the sounds that will enhance the experience for the listener thereby reinforcing the Director's intention for a given scene. The artist must decide what artificial sound captures the essence of both the sound and action depicted in the scene and then manually record them, edit, and layer and synchronise them one by one with the action happening on screen. Apart from a tedious task, it also involves the artistic input and imagination of the folley artist. AutoFoley is a fully-automated deep-learning tool that can be used to synthesise a representative audio track for videos. It can be used in applications where there is either no corresponding audio file associated with the video, or in cases where there is a need to identify critical scenarios and provide a synthesised, reinforced sound track. An important performance criterion of the synthesised sound track is to be time-synchronised with the input video, which provides for a realistic and believable portrayal of the synthesised sound. Unlike existing sound prediction and generation architectures, this algorithm is capable of precise recognition of actions as well as inter-frame relations in fast moving video clips by incorporating an interpolation technique and Temporal Relational Networks (TRN). A robust multi-scale Recurrent Neural Network (RNN) associated with a Convolutional Neural Network (CNN) is employed for a better understanding of the intricate input-to-output associations over time. To better evaluate AutoFoley, the authors created and introduced a large-scale audio-video dataset containing a variety of sounds frequently used as Foley-effects in movies. Their experiments show that the synthesised sounds are realistically portrayed with

accurate temporal synchronisation of the associated visual inputs. Human qualitative testing of AutoFoley show over 73% of the test subjects considered the generated sound track as original, which is a noteworthy improvement in cross-modal research in sound synthesis. This is a groundbreaking tool in the hands of a human editor, since it eliminates the time of the production of foley which will cut down time and cost from the production of many films, especially the ones heavily based on foley for the editor to work on. (e.g. animation or action films)

Moving on from the use of audio and speech synthesis, we can explore more examples of Phase 3 focusing more on "classic" editing scenarios and their implementations. More precisely, we'll examine examples of Automatic Creation of Edited Scenes Using Multicamera and Single Camera Originated Footage. This example seems as an optimal "end state" for the application of AI and ML for automatic content creation. The combination of Phases 1–3 culminates in the intelligent, coherent, and automatic creation of content. The study is ongoing and has resulted in demonstrable prototypes for two very common use cases.

- Multicamera originated footage that must be edited into a coherent final scene.
- Single-camera, multiple angle, and multiple-take originated footage that must be edited into a coherent final scene.

Footage must be analysed to determine what is of primary interest, at what moment in time, and when a change is made. With multicamera footage, there are some number of cameras, angles, and focal lengths—a myriad of possibilities in how the final scene can be constructed. Research by Arev et al.[28] is particularly noteworthy. Four consumer cameras are used to simultaneously capture the footage of the same event from different angles. Through the use of algorithms developed by examining what is considered to be the principal point of interest and based on acceptable and known rules of cinematography, it is

possible to create working models that automate the construction of scenes based on this type of footage. For example, an algorithmic model is created, based on the 180° rule of cinematography so that the action is always presented in its proper spatial relationship. Furthermore, an algorithm was created to avoid the creation of “jump cuts,” where cutting to a camera at an angle too close to the immediately previous angle creates the appearance of a jump in action. Footage from social cameras contains an intimate, personalised view that reflects the part of an event that was of importance to the camera operator (or wearer). They leveraged the insight that social cameras share the focus of attention of the people carrying them and they used this insight to determine where the important "content" in a scene is taking place, and use it in conjunction with the aforementioned cinematographic guidelines to select which cameras to cut to and to determine the timing of those cuts. They demonstrated cuts of the videos in various styles and lengths for a number of scenarios, including sports games, street performances, family activities, and social get-togethers. They evaluated the results through an in-depth analysis of the cuts in the resulting videos and through comparison with videos produced by a professional editor and existing commercial solutions. Finally, combining these algorithms for the 180° rule, jump-cut avoidance, and 3D camera motion estimation, a model was created for the automatic editing of multicamera footage. The results are not satisfactory for every scenario. Some scenes were highly challenging, and there may be cases where the semantics of the scene was too complex to be so simply captured. Audio could also be a significant cue in determining the location of content in the scene, so it was a factor to be considered. Lastly, as a general remark they note that the algorithm can also be used to assist professional editors in their task of editing large amounts of footage by providing several possible different movies to choose from. It would be interesting to build an interface for such a semi-automatic editing tool, thus

admitting that the algorithm cannot compete with a professional editor being a creator of its own.

The second common case, the Automatic Creation of Edited Scenes Using Single Camera Originated Footage has also been extensively studied. Intricacies are encountered when a single camera, multiple angles, and multiple take footage are the raw content inputs. Consider the variables for a two-person dialog-centric scene: What character is on-screen? What angle is used for that character? When should a cut be made to the second character?

A study by Leake et al.[29] at Stanford Research and Adobe Research, is making use of idiom-based editing and applies accepted storytelling rules and motion picture editing concepts to automated content creation.

Automatic video editing is an artistic process involving at least the steps of selecting the most valuable footage from the points of view of visual quality and the importance of the action filmed; and cutting the footage into a brief and coherent visual story that would be interesting to watch is implemented in a purely data-driven manner. The authors describe a system that is capable of learning the editing style from samples extracted from the content created by professional editors, including motion picture masterpieces, and of applying this data-driven style to cut non-professional videos with the ability to mimic the individual style of selected reference samples. Visual semantic and aesthetic features are extracted by an ImageNet-trained convolutional neural network, and the editing controller can be trained by an imitation learning algorithm or reinforcement learning algorithm. As a result, during the test the controller showed signs of observing basic cinematography editing rules learned from the corpus of motion pictures masterpieces. The loss function developed for learning approaches can be efficiently applied in a global optimisation setting of the automatic video editing problem using dynamic programming.

The idioms that are used to define the editing parameters are:

- Avoid jump cuts: Avoid transitions between clips that show the same visible speakers to prevent jarring transitions
- Change zoom gradually: Avoid large changes in the zoom level that can disorient viewers and instead change zoom levels slowly.
- Emphasise character: Do not cut away from shots that focus on an important character unless another character has a long line. This focuses the audience's attention on the more important character.
- Intensify emotion: Use close-ups for particularly emotional lines to provide more detail in the performer's face.
- Mirror position: Select clips for one performer that most closely mirror the screen position of the other performer to create a back-and-forth dynamic for a two person conversation.
- Peaks and valleys: Zoom in for more emotional lines and zoom out for less emotional lines to allow the audience to see more detail in the performer' faces for emotional lines.
- Performance fast/slow: Select shorter (longer) performances of a line to control the pacing of the scene.
- Performance loud/quiet: Select louder (quieter) performances of a line to control the volume of the scene.
- Short lines: Avoid cutting to a new shot for only a short amount of time to prevent rapid, successive cuts that can be jarring.
- Speaker visible: Show the face of the speaking character on screen to help the audience keep track of which character is speak- ing and understand the progression of the conversation.
- Start wide: Start with the widest shot possible to establish the scene (i.e., start with establishing shot) and show the relation- ship between performers and the surroundings.

- Zoom consistent: Maintain a consistent zoom level throughout the scene to create a sense of balance between the performers.
- Zoom in/out: Specify a preference for zooming in (out) throughout a scene to reveal more (less) detail in performers' faces and create more (less) intimacy.

In one such research prototype, the audio, video, and text analysis of raw content result in the extraction of a variety of labels from takes and the script. It then segments each take into separate clips for each line of dialog and aligns them with the text of the script. This correlation between the input script and the lines of the character spoken dialog is already a familiar tool of editors (Avid's Script Sync™). Next, in order for a model to be created which can automatically identify a speaker as well as when a speaker is talking, facial analysis and motion tracking technology are employed.

Idioms that correspond to accepted and time-honoured picture editing rules are used to automatically construct a scene. Furthermore, idiomatic expressions derived from the ruleset include: avoiding jump cuts, emphasising certain characters, and pacing characteristics. The idioms are decided by the creators of the tool, thus vary according to the implementation. The result of this model is the automatic creation of a scene that can be manipulated in an infinite manner, just as if the scene were being edited by a human. A graphical user interface that presents these functions is also available, allowing for basic manipulations, based on the idioms mentioned above. Although the results are not compared with the work of a professional editor, the author acknowledges one serious limitation of the proposed method, which is the linear narration. It can only work for an ordered sequence of footage only, and reordering video takes or inserting B-roll shots is not possible in a general way. It is more oriented as an interactive gamification of the video editing process in the form of trial and error in setting

various parameters that could allow non-professional users to achieve quite pleasing results without requiring technical and artistic skills.

Combining Phases 1–3 technologies can result in the creation of content that is a complete fabrication. Consider the following predicament: after editing a scene, it is determined that a shot of an actor is needed. Unfortunately, the shot was never produced and different focal lengths are required. Furthermore, it is necessary that the actor say certain lines, but the actual audio was never produced. Last, to complicate everything— the actor has since passed away.

Based on the research done by Suwajanakorn and Seitz[30] (University of Washington), a synthetic (in this context—entirely fabricated) clip with accurate lip synchronisation was created. Using multiple hours of audio of President Obama, a "recurrent neural network learns the mapping from raw audio features to mouth shapes." With a database of mouth shapes associated with time instances, mouth textures were synthesised and then composited with 3D matching to change what he appears to be saying. The result is that synthetic photorealistic shots can be created.

On the same premise, this vary same technology is reported of being used to create fake news or abusive spam on social media. Deepfake technologies can also create realistic fake videos by replacing some parts of the media with synthetic content. For example, substituting someone's face whilst hair, body and action remain the same. More recently, DeepFaceLab[31] provided a state of the art tool for face replacement; however manual editing is still required in order to create the most natural appearance. Whole body movements have been generated via learning from a source video to synthesise the positions of arms, legs and body of the target in[32]. Despite rapid progress in this area, the creation of perfectly natural figures remains challenging; for example deepfake faces often do not blink naturally. Deepfake techniques have been widely used to create pornographic images of celebrities, to cause political distress or social

unrest, for purposes of blackmail and to announce fake terrorism events or other disasters. This has resulted in several countries banning non-consensual deepfake content. To counter these often malicious attacks, a number of approaches have been reported and introduced to detect fake digital content[33].

On the contrary, the same technology used to create deep fakes, is one of the most groundbreaking tools that can be used for today's movie production, affecting straightforwardly how an editor is dealing with the performances of the actors. Using face or mouth swapping technologies, in association with speech synthesis, the editor is allowed to alternate any take or shot of his choice and eliminate any "mistakes" from the shooting or the performance, rendering all takes as usable for the edit. This procedure is currently called "vubbing", a combination of the words video+dubbing, coined by Flawless AI™.

The commercialisation of AI-assisted dubbing for TV and films hints at bigger possibilities with digitally captured acting performances. Hollywood has already used computer-aided visual effects to make veteran actors look decades younger, enable action stars such as Arnold Schwarzenegger and Will Smith to battle digital doubles of themselves, and virtually resurrect dead actors for new Star Wars films or TV commercials. Now AI that can learn the idiosyncrasies of each actor's voice and facial expressions is making it easier to dub films and TV shows in new languages while preserving the acting nuances and voices of the original-language performances.

The London-based startup Flawless AI has partnered with researchers at the Max Planck Institute for Informatics in Germany[34] to commercialise technology that can digitally capture actors' performances from 2D film or TV image frames and transform those into a 3D computer model. After training AI to learn specific actors' vocal and facial performances, the startup can generate modified versions of the original performance that change the actor's voice and facial expression to fit an entirely different language.

“It's actually a pixel perfect 3D representation of the head of each of the actors,” says Nick Lynes, co-CEO and founder of Flawless AI. “And because of that pixel perfect frame, it represents every single phenomena and idiosyncratic style possible that the actor does, because it doesn't take much before the AI has understood all of the idiosyncrasies.”

The AI-generated dubbing performances still require some manual touch-ups by human visual effects artists, and Lynes expects that to be the case for the foreseeable future. But the impressive end results showcased in a Flawless AI demo reel include Tom Cruise and Jack Nicholson confronting each other in fluent French in the 1992 film “A Few Good Men,” Robert DeNiro speaking German in the 2015 film “Heist,” and Tom Hank’s titular “Forrest Gump” character crying over Jenny’s grave while speaking German, Spanish, and Japanese in the 1994 film.

Several other companies have been using AI to dub movies, TV shows, advertisements, and other content in new languages while retaining the original voices, including Israeli companies such as Deepdub and Canny AI. But Flawless AI’s approach goes beyond just redoing the audio by reshaping the actor’s mouth movements and facial expressions to suit the new language dub. It’s likely one of the very first efforts to commercialise such technology beyond what has been demonstrated in academic research papers.

The company can currently perform multiple dubs within the usual production time-frame for a film or TV show, or it can go back and do dubbing for an older film within six to 10 weeks. It has also figured out how to reduce the amount of training time and data necessary for its AI to learn all the performance nuances of each actor. The team initially used all the available raw, unedited film footage from each film or TV production to train the AI, but has since figured out how to more efficiently train its AI using just a small fraction of such footage.

Digitally modifying a person's mouth movements to fit entirely different words is something that can also be seen in AI-assisted "deepfake" videos. But Flawless AI's technology delivers more realistic and natural modified performances than the typical deepfake video, which is important given that the end results may need to be shown on a standard movie theatre screen.

But the Flawless AI approach to digitising actor performances may have even bigger implications beyond enabling seamless dubbing of the newest films and TV shows in multiple languages. The company's technology not only digitally captures the actor's facial expressions and mouth movements, but also trains the AI on patterns in the walking gait and body movements of the actors.

That raises the possibility of using the actors' digital doubles to modify certain movie scenes in ways that fix mistakes or better conform to a director's vision instead of getting the actors and crew back together for expensive reshoots. Combining this with the newest advancements on moving image Frame Interpolation[35], it's not hard to imagine how such technology could make it much easier for a Hollywood studio to change something such as removing Henry Cavill's contractually-obligated moustache from the face of Superman during reshoots for the 2017 film "Justice League."

The Flawless AI approach also digitally captures the entire environment and surroundings of an actor in a given scene. That could make it easier for algorithms to remove mistakes such as the infamous coffee cup that appears in the last season of the HBO medieval fantasy series "Game of Thrones," as Lynes pointed out. "Sometimes in the live action world, the real world, it's hard to make sure everything goes in the way that you want it to go to be able to tell your story exactly as you wanted to tell it," Lynes says. "So having AI [enabled] visual effects gives me functions in the future that are not possible to do with the standard visual effects process."

Part of this toolset has already made its appearance in commercial DNLEs. Editors and their assistants can already benefit from its potential and use locally the advantages of the above mentioned technologies. The application of AI and ML will have profound effects on the content creation process. The automated creation of content is necessary to address the ever-increasing variety of content and venues that must be served. One could understand that these projections might be received with skepticism, particularly with respect to automatic content fabrication. However, based on the current and predicted technology development trajectories, the most logical conclusion is that the content creation process will be very much impacted and influenced by these three phases.

## 4. Conclusion

Throughout this thesis we have reviewed on the successes of AI in supporting and enhancing processes where there is good availability of data as a basis for machine learning. We have seen that AI-based techniques work very well when they are used as tools for information extraction, analysis and enhancement. Deep learning methods that characterise data from low-level features and connected these to extract semantic meaning are well suited to these applications. AI can thus be used with success, to perform tasks that are too difficult for humans or are too time-consuming, such as searching through a large database and examining its data to draw conclusions. Post production workflows will therefore see increased use of AI, including enhanced tools for denoising, colorisation, segmentation, rendering and tracking. Motion and volumetric capture methods will benefit from enhanced parameter selection and rendering tools. Virtual production methods and games technologies will see greater convergence and increased reliance on AI methodologies.

In all the above examples, AI tools will not be used in isolation as a simple black box solution. Instead, they must be designed as part of the associated workflow and incorporate a feedback framework with the human in the loop. For the foreseeable future, humans will need to check the outputs from AI systems, make critical decision, and feedback ‘faults’ that will be used to adjust the model. In addition, the interactions between audiences or users and machines are likely to become increasingly common. For example, AI could help to create characters that learn context in location-based story-telling and begin to understand the audience and adapt according to interactions.

Currently, the most effective AI algorithms still rely on supervised learning, where ground truth data readily exists or where humans have labelled the dataset prior to using it for training the model. In contrast, truly creative processes do not have predefined outcomes that can simply be classed as good

or bad. Although many may follow contemporary trends or be in some way derivative, based on known audience preferences, there is no obvious way of measuring the quality of the result in advance. Creativity almost always involves combining ideas, often in an abstract yet coherent way, from different domains or multiple experiences, driven by curiosity and experimentation. Hence, labelling of data for these applications is not straightforward or even possible in many cases. This leads to difficulties in using current ML technologies.

In the context of creating a new artwork, generating low-level features from semantics is a one-to-many relationship, leading to inconsistencies between outputs. For example, when asking a group of artists to draw a cat, the results will all differ in color, shape, size, context and pose. Results of the creative process are thus unlikely to be structured, and hence may not be suitable for use with ML methods. Unfortunately, the currently used models, are not yet sufficiently robust to consistently create results that are realistic or valuable. It is clear that significant additional work is needed to extract significant value from AI in this area.

The research and development of AI-based solutions will not cease to continue with an extreme pace. The more it develops and the new algorithms deployed are becoming more effective, requiring just a fraction of the time and processing power needed today, AI will be attracting major investments from governments and large international organisations alongside everyday applications. ML algorithms will be the primary driver for most AI systems in the near future and AI solutions will impact profoundly a broader range of sectors, including the media industry. The pacing of AI research advancements has been profited from our recent ability to generate, access and store massive amounts of data effectively, and on advances in graphics processing architectures and parallel computing to process these massive amounts of data. In the future, new

computational solutions such as the famous quantum computing, will likely play a groundbreaking role in this field[36].

In order to produce an original work, such as editing, it would be beneficial to support increased diversity and context when training AI systems. The quality of the solution in such cases is difficult to define and will inevitably depend on audience preferences and the inevitable comparison to professionally created content. High-dimensional datasets, carefully curated, that can represent some of these characteristics will therefore be needed. Furthermore, the parameters that dictate the convergence of the AI's internal algorithms must reflect perceptions rather than simple mathematical models. Research into such parameters that better reflect human perception of performance and quality is therefore an area for further research.

Since ML-based AI algorithms are data-driven, the curating process of how to select and prepare datasets for creative applications will be key to future developments. Defining, cleaning and organising bias-free data for creative applications are not straightforward tasks. The task of data collection and labelling can be highly resource intensive, and labelling services are starting to have a rising demand nowadays.

As the amount of unlabelled data grows dramatically, unsupervised or self supervised ML algorithms are prime candidates for underpinning future advancements in the next generation of ML. There exist techniques that will greatly benefit the applications of Phase 1, and lead to better end results and implementation in the current workflows. Unfortunately the same cannon be stated for the majority of applications of Phase 2 and especially not for Phase 3.

It is clear that current AI methods do not mimic the human brain, or even parts of it. The data driven learning approach with error backpropagation is not apparent in human learning. Humans learn in complex ways that combine genetics, experience and prediction-failure reinforcement. A typical example is

the editing procedure of a documentary, where a multitude of parameters should be taken into consideration in order to create a cohesive result, set aside follow the vision of the director. The amount of parameters an AI editing application should deal with prohibits any viable result.

This thesis has presented a comprehensive review of current AI technologies and their applications, specifically in the context of the creative industries related to editing workflows. We have seen that ML-based AI has advanced the state of the art across a range of creative applications including content creation, information analysis, content enhancement, information extraction, information enhancement and data compression. ML–AI methods are data driven and benefit from recent advances in computational hardware and the availability of huge amounts of data for training – particularly sound and video data.

We have differentiated throughout between the use of ML–AI as a creative tool and its potential as a creator in its own right. We foresee, in the near future, that AI will be adopted much more widely as a tool or collaborative assistant for creativity, supporting acquisition, production, post-production, delivery and interactivity. The concurrent advances in computing power, storage capacities and communication technologies will support the embedding of AI processing within and at the edge of the network. In contrast, we observe that, despite recent advances, significant challenges remain for AI as the sole generator of original work. ML–AI works well when there are clearly defined problems that do not depend on external context or require long chains of inference or reasoning in decision making. It also benefits significantly from large amounts of diverse and unbiased data for training. Hence, the likelihood of AI (or its developers) winning awards for creative works in competition with human creatives may be some way off. We therefore conclude that, for creative applications, technological developments will, for some time yet, remain human-centric.

Designed to augment, rather than replace, human creativity. As AI methods begin to pervade the creative sector, developers and deployers must however continue to build trust; technological advances must go hand-in-hand with a greater understanding of ethical issues, data bias and wider social impact.

## Bibliography

[1] NSTC, “Preparing for the future of artificial intelligence,” 2016, https: //obamawhitehouse.archives.gov/sites/default/files/whitehouse files/microsites/ostp/ NSTC/preparing for the future of ai.pdf, accessed 10-04-2020.

[2] D. W. Hall and J. Pesenti, “Growing the artificial intelligence industry in the UK,” 2018, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/ attachment data/file/652097/Growing the artificial intelligence industry in the UK.pdf.

[3] Gentile, A. (2022). How AI is Changing TV Production for the Better. Retrieved 23 August 2022, from https://www.smpte.org/blog/how-ai-changing-tv-production-better

[4] J. Rowe and D. Partridge, “Creativity: a survey of AI approaches,” Artif Intell Rev, vol. 7, pp. 43–70, 1993.

[5] https://www.pfeifferreport.com/wp-content/uploads/2018/11/Creativity and AI Report INT.pdf

[6] B. H. Lange, "Artificial Intelligence and SMPTE," in SMPTE Motion Imaging Journal, vol. 130, no. 3, pp. 6-6, April 2021, doi: 10.5594/JMI.2021.3062889.

[7] S. C. Bilow, "AI and Machine Learning," in SMPTE Motion Imaging Journal, vol. 130, no. 3, pp. 10-11, April 2021, doi: 10.5594/JMI.2021.3057688.

[8] P. J. Hayes and L. Morgenstern, “On John McCarthy’s 80th Birthday, in Honor of His Contributions,” *AI Magazine. Association for the Advancement of Artificial Intelligence*, 2007.

[9] B. Devlin, "AI and ML," in SMPTE Motion Imaging Journal, vol. 130, no. 3, pp. 8-8, April 2021, doi: 10.5594/JMI.2021.3064082.

[10] A. Turing, “Computing Machinery and Intelligence,” *Mind*, LIX (238):433–460, Oct. 1950.

[11] R. Welsh, "Defining Artificial Intelligence," in SMPTE Motion Imaging Journal, vol. 128, no. 1, pp. 26-32, Jan.-Feb. 2019, doi: 10.5594/JMI.2018.2880366.

[12] “Preliminary study on the ethics of artificial intelligence,” 2019, SHS/COMEST/EXTWG-ETHICS-AI/2019/1: https://unesdoc.unesco.org/ark/48223/ pf0000367823.

[13] Siwicki, Bill. “How AI Bias Happens – and How to Eliminate It.” *Healthcare IT News*, 30 Nov. 2021, www.healthcareitnews.com/news/how-ai-bias-happens-and-how-eliminate-it.

[14] Samuel, Sigal. “AI Bias: Why Fair Artificial Intelligence Is so Hard to Make.” *Vox*, 19 Apr. 2022, www.vox.com/future-perfect/22916602/ai-bias-fairness-tradeoffs-artificial-intelligence.

[15] Abid, A., Farooqi, M. & Zou, J. Large language models associate Muslims with violence. *Nat Mach Intell* **3,** 461–463 (2021). https://doi.org/10.1038/s42256-021-00359-2

[16] Solaiman, Irene & Dennison, Christy. (2021). Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets.

[17] ---. “AI Bias: Why Fair Artificial Intelligence Is so Hard to Make.” *Vox*, 19 Apr. 2022, www.vox.com/future-perfect/22916602/ai-bias-fairness-tradeoffs-artificial-intelligence.

[18] V. Dignum, “Ethics in artificial intelligence: Introduction to the special issue,” Ethics Inf Technol, pp. 1–3, 2018.

[19] T. Ohanian, "How Artificial Intelligence and Machine Learning Will Change Content Creation Methodologies," SMPTE 2017 Annual Technical Conference and Exhibition, 2017, pp. 1-15, doi: 10.5594/M001794.

[20] Anantrasirichai, N., Bull, D. Artificial intelligence in the creative industries: a review. *Artif Intell Rev* **55,** 589–656 (2022). https://doi.org/10.1007/s10462-021-10039-7

[21] J. Priestly, “Wimbledon to Use IBM’s Watson for Highlights, Analytics,” *TV Technology*, July 2017.

[22] A. Bera, "Artificial Intelligence: Transforming the Live Sports Landscape," in SMPTE Motion Imaging Journal, vol. 130, no. 3, pp. 28-34, April 2021, doi: 10.5594/JMI.2021.3060823.

[23] H. Zen, K. Tokuda, and A. W. Black, “Statistical Parametric Speech Synthesis,” *Speech Comm.*, 51(11):1039–1064, Nov. 2009.

[24] K. Kurihara, N. Seiyama, T. Kumano, T. Fukaya, K. Saito and S. Suzuki, "“AI News Anchor” With Deep Learning-Based Speech Synthesis," in SMPTE Motion Imaging Journal, vol. 130, no. 3, pp. 19-27, April 2021, doi: 10.5594/JMI.2021.3057703.

[25] K. Kurihara et al., "Automatic Generation of Audio Descriptions for Sports Programs," in SMPTE Motion Imaging Journal, vol. 128, no. 1, pp. 41-47, Jan.-Feb. 2019, doi: 10.5594/JMI.2018.2879261.

[26] A. Malik and H. Nguyen, "Exploring Automated Voice Casting for Content Localization Using Deep Learning," in SMPTE Motion Imaging Journal, vol. 130, no. 3, pp. 12-18, April 2021, doi: 10.5594/JMI.2021.3057695.

[27] Ghose, Sanchita & Prevost, John. (2020). AutoFoley: Artificial Synthesis of Synchronized Sound Tracks for Silent Videos With Deep Learning. IEEE Transactions on Multimedia. PP. 1-1. 10.1109/TMM.2020.3005033.

[28] Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. 2014. Automatic editing of footage from multiple social cameras. ACM Trans. Graph. 33, 4, Article 81 (July 2014), 11 pages. https://doi.org/10.1145/2601097.2601198

[29] M. Leake, A. Davis, A.Truong, and M. Agrawala, “Computational Video Editing for Dialog-Driven Scenes,” *ACM Trans. Graph.*, 36(4):130, July 2017.

[30] S. Suwajanakorn and S. M. Seitz, “Synthesizing Obama: Learning Lip Sync from Audio,” *ACM Trans. Graph.*, 36(4):95, July 2017.

[31] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Um´e, M. Dpfks, C. S. Facen- heim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang, “Deepfacelab: A simple, flexible and extensible face swapping framework,” arXiv preprint arXiv:2005.05535v4, 2020.

[32] C. Chan, S. Ginosar, T. Zhou, and A. Efros, “Everybody dance now,” in IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5932–5941.

[33] Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[34] Hsu, Jeremy. “AI Modifies Actor Performances for Flawless Dubbing.” *IEEE Spectrum*, 23 July 2021, spectrum.ieee.org/ai-modifies-actor-performances-for-flawless-dubbing.

[35] A. Kokaram, D. Singh and S. Robinson, "Moving Image Frame Interpolation: Neural Networks and Classical Toolsets Compared," in SMPTE Motion Imaging Journal, vol. 130, no. 4, pp. 1-11, May 2021, doi: 10.5594/JMI.2021.3067960.

[36] J. Welser, J. W. Pitera, and C. Goldberg, “Future computing hardware for AI,” in IEEE International Electron Devices Meeting (IEDM), 2018, pp. 1.3.1–1.3.6.