Technische Hochschule Ingolstadt

Bachelor Thesis

in the Degree Program Bachelor User Experience Design (UXD)

Faculty Computer Science

Ethical and Societal Implications of Generative AI-Models

First and last name: Felix Reichwein

Issuing date: 04.10.2023

Date of hand-in: 26.02.2024

First Supervisor: Prof. Dr. Matthias Uhl

Second Supervisor: Prof. Dr. Andreas Riener

Declaration/Affidavit

I hereby declare that I have composed this work independently, have not previously submitted it for examination purposes elsewhere, have not used any sources or aids other than those indicated, and have duly acknowledged all direct and indirect quotations as such.

Ingolstadt, ____________________

Felix Reichwein

Declaration on the Use of Artificial Intelligence

This bachelor's thesis was written independently. The Large Language Model (LLM) "ChatGPT" was only used for support in the creation of the text, specifically for assistance with phrasing and the identification of synonyms. The core content and the conception of the work are solely attributable to the author. No external writing services or third parties were engaged to compose the text.

Ingolstadt, ____________________

Felix Reichwein

Abstract

This thesis examines the ethical and societal implications of generative AI models, focusing on their associated risks, but also exploring the potential benefits. It synthesizes existing research to outline the primary concerns and opportunities that generative AI presents, with an emphasis on the need for its responsible development and deployment. The analysis includes an exploration of the increasing difficulty in distinguishing between AI-generated and human-created content, highlighting advancements in AI that challenge current notions of authenticity and trust.

The work further discusses a variety of broader societal changes and risks that may emerge with generative AI, advocating for the necessity of regulatory interventions to mitigate potential negative outcomes. It assesses and critically investigates the current regulatory landscape and potential regulation strategies, noting the lack of comprehensive and mandatory frameworks to address the implications of generative AI adequately.

The critical role of collaborative efforts among stakeholders in navigating the ethical landscape of generative AI is highlighted. It concludes that addressing the examined challenges requires a multidisciplinary approach, integrating insights from an array of fields, including technology, ethics, and policy to ensure that generative AI advances in a way that aligns with societal values and norms.

Contents

Abbreviations

AI	*Artificial Intelligence*
CEO	*Chief Executive Officer*
GDP	*Gross domestic product*
UBI	*Universal Basic Income*
XAI	*Explainable Artificial Intelligence*

1 Introduction

“I've always thought of AI as the most profound technology humanity is working on, more profound than fire, electricity, or anything else we have done in the past” (Sundar Pichai, 2023). This is how Sundar Pichai, CEO of Google and its parent company Alphabet, describes AI in an interview in 2023. Bold statements like this underline both the significance and the ambiguity of AI. It is a technology that has rapidly integrated into various aspects of daily life, from smart homes and healthcare to transportation and entertainment (Gozalo-Brizuela & Garrido-Merchán, 2023).

Particularly, generative AI has recently experienced a dramatic increase in public and academic interest, leading to the emergence of various applications and discussions about its capabilities and implications. The widespread attention towards generative AI has induced a rush to implement AI solutions across different sectors, initiating debates on their potential benefits and challenges (LaGrandeur, 2023).

This thesis aims to explore the ethical and societal implications of generative AI, navigating through the optimism and concerns surrounding its development and application. By combining existing research, it seeks to provide a comprehensive overview of the key issues associated with generative AI, with a focus on critically analyzing its potential impacts and providing insights into the responsible development and application of generative AI technologies.

Regarding the structure of the thesis, the first paragraphs aim to provide a fundamental understanding of AI and generative AI by covering definitions, historical evolution, as well as current advancements and applications. This foundation supports the more in-depth examination of technical aspects.

Following the initial chapter, the thesis investigates the current state of distinguishability between AI-generated content and human-created content across various modalities, explaining the importance of this issue.

The third chapter extensively examines the ethical and societal implications of generative AI. While acknowledging the significant positive potential of these technologies, the focus is primarily on addressing the concerns and negative implications, as the need for proactive risk mitigation is especially prominent.

The subsequent sections assess existing and proposed strategies for regulating and controlling generative AI and its associated risks. This analysis critically evaluates the effectiveness of current mitigation approaches.

Concluding the thesis, the final chapters provide a look on the future of this technology, drawing conclusions from the discussions presented throughout the thesis and acknowledging the research limitations encountered.

1.1 Definition of AI

The term 'AI' is often misinterpreted, as many misconceptions surround the topic, potentially due to the vagueness of the term (Emmert-Streib et al., 2020). It is often used interchangeably with machine learning, neural networks, and other related technologies, even though they are not quite synonymous and differ significantly (IBM Data and AI Team, 2023).

To clarify this often-misunderstood field of AI, we can start by examining its definitions. The European Commission describes it as systems that *“display intelligent behaviour by analysing their environment and taking actions—with some degree of autonomy—to achieve specific goals*” (European Comission, 2018, p. 1). Another common definition, quite similar in essence, describes AI as *“a system’s ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation”* (Haenlein & Kaplan, 2019, p. 1)

These definitions underscore key elements such as intelligent behavior, environmental analysis, autonomy, and goal-oriented actions. However, defining AI is not a straightforward task, as even for human intelligence there isn’t one definition that is agreed upon. The field of AI has historically been approached from several different angles (S. J. Russell et al., 2022). With that in mind, the different kinds of artificial intelligences are now further examined.

1.2 Narrow and General AI

In general, there are two types of artificial intelligence: General AI and narrow AI. “*A general AI system is intended to be a system that can perform most activities that humans can do. Narrow AI systems are instead systems that can perform one or few specific tasks. Currently deployed AI systems are examples of narrow AI”* (European Comission, 2018, p. 7). The development of general AI is a high priority in the field and recent research does show sparks of general AI being observed as of early 2023 (Bubeck et al., 2023). Therefore, the near future could show AI becoming more generally applicable.

Most narrow AI systems are typically trained on thoroughly selected datasets to improve abilities within a specific task range. In contrast, the approaches for general AI are more diverse. This diversity reflects varying perspectives on the capabilities a general AI system should possess. These perspectives on actual intelligence in machines can be categorized based on the dimensions ‘human vs. rational’ and ‘thought vs. behavior’. These rather abstract divisions create four possible combinations of AI capability: 1) Acting humanly; 2) Thinking humanly; 3) Thinking rationally; 4) Acting rationally. Each of them has special requirements, as well as disadvantages and benefits, therefore the approaches can be quite diverse, depending on the goal. Generally, by now ‘acting rationally’ is mostly accepted as the standard model for actual machine intelligence. (S. J. Russell et al., 2022)

These combinations "*can be classified into analytical, human-inspired, and humanized AI depending on the types of intelligence it exhibits (cognitive, emotional, and social intelligence)*" (Haenlein & Kaplan, 2019, p. 2).

1.3 Evolution of AI

The evolution of AI as it is known today began in the early 20th century when the foundations were laid with the introduction of the concept machine intelligence. By the mid-1950s, the term "Artificial Intelligence" was invented, signifying the advancement of dedicated AI research. This period of optimism, often referred to as the AI Summer, saw the development of tools that could simulate human conversation and problem-solving. However, by the early 1970s, this perspective faced criticism as only little results were achieved despite high investments (Haenlein & Kaplan, 2019).

In the contemporary time, Deep Learning emerged. Deep Learning systems are based on neural networks and attempt to simulate the behavior of the human brain. It is another milestone of AI, driving numerous advancements. While models based on Deep Learning mechanisms have shown promising results, their internal mechanisms are more complex and less easy to trace back, leading to challenges regarding transparency and understanding (Ongsulee, 2017). With the availability of vast datasets and improved computing power, AI has found use cases in diverse topics, from image recognition to autonomous vehicles and meanwhile even into many private households. The introduction of ChatGPT (OpenAI, 2022), an advanced conversational AI, has significantly contributed to the mainstream adoption of artificial intelligence technologies (Grassini, 2023). In certain areas, AI tools have even demonstrated the capability to outperform human experts, for example in specific medical diagnostics (K. Cao et al., 2023; Shen et al., 2019) and currently deployed models are frequently breaking previous records (Taulli, 2023). Throughout its history, AI's trajectory has been shaped by technological innovations, with periods of swift progress and stagnation (Haenlein & Kaplan, 2019).

1.4 Introduction to Generative AI

The core concept of generative AI involves creating completely new content by understanding and applying the patterns and distributions in existing data. Unlike conventional AI systems that focus on specific tasks and follow predefined rules, Generative AI possesses the ability to generate creative outputs like text, music, images, and beyond, simulating aspects of human creativity (Nalini et al., 2023).

This unprecedented possibility to generate entirely new content, coupled with the democratization of this power due to the public availability of these models, are some of the reasons why AI has recently become so pervasive. Particularly, the natural-language-based interaction with many contemporary products has transitioned AI from a subject exclusive to specialists to something anyone can utilize.

1.5 Generative AI Applications

A few years ago, the area of generative AI was primarily limited to textual outputs. However, recent advancements have diversified its modalities, now containing a broad spectrum of possible input and output modalities. Even though there are many possible forms of input, text remains the most common, mostly due to the ability to communicate complex concepts with minimal effort using natural language. Text output is equally versatile. In general, the modalities can be categorized under several umbrella terms, including text, sound, image, video, code/software, and 3D, each with further subdivisions into more specific applications (Y. Cao et al., 2023; Gozalo-Brizuela & Garrido-Merchán, 2023). The various output modalities are now explored in greater depth subsequenty.

Text output:

For instance, speech can be transcribed and AI models can interpret and describe other modalities like images and music. Notably, conversational AIs, such as ChatGPT, remain a prominent application. A few applications of this conversational AI include generating texts in scientific language, mimicking the style of specific authors and providing medical advice, albeit there are many more use-cases (Gozalo-Brizuela & Garrido-Merchán, 2023; Nalini et al., 2023).

Image output:

The field of digital images is categorized into vector- and pixel-based pictures and thus also the image generation by generative AI. Both categories allow for the creation of images from scratch or the editing of existing images in various ways (Gozalo-Brizuela & Garrido-Merchán, 2023). While providing some advantages regarding editability, a limitation of the vector-based generation is the incapability to create realistic images. The pixel-based approach, more common and versatile, spans artistic, stylized, and realistic forms. A notable application involves extending real photographs authentically when provided with a reference image series (L. Tang et al., 2023).

Sound output:

In the sound domain, speech, with the possibility of communication, is a primary focus due to its wide range of applications, including interaction, entertainment, and accessibility. Generative AI can edit, recognize, and clone voices (Y. Cao et al., 2023; Zhang et al., 2019). Music generation is another significant application, with inputs varying from text and existing music to dance recordings, often incorporating voice generation. Whilst the most relevant, the scope of generative sound extends beyond speech and music, encompassing a variety of other auditory forms (Gozalo-Brizuela & Garrido-Merchán, 2023).

Video output:

Video content generated by AI can also be either realistic or stylized. Either one of them, but especially the quality of realistic AI-generated video content has seen substantial improvements recently (Ho et al., 2022). Generative AI moreover enables video editing, including the creation of deepfakes (face replacement) and the insertion or modification of objects in videos (Chadha et al., 2021; Westerlund, 2019). Emerging applications facilitate easy translation dubbing of speech with corresponding tonality duplication and lip movement adaptation (Gozalo-Brizuela & Garrido-Merchán, 2023).

3D output:

Generative AI can produce 3D models and landscapes from inputs such as texts, static images, or videos. The styles can vary from stylized models to 3D models that mimic reality (Gozalo-Brizuela & Garrido-Merchán, 2023). New emergent techniques enable the real-time creation of realistic 3D models using simple tools (Kerbl et al., 2023).

Code output/software:

The software industry has already experienced significant changes due to generative AI. Models capable of generating code have quickly altered traditional coding practices. These models can create, transform, complete and translate code across different programming languages or generate web pages and applications from an image of the desired outcome. With the models' capacity to execute steps and programs themselves, new possibilities emerge. They enable automation of software processes, cybersecurity testing, and facilitate more autonomous robotic operations (Gozalo-Brizuela & Garrido-Merchán, 2023).

Multimodal:

Multimodality in generative AI refers to the integration of multiple modalities, enabling the use of various inputs and outputs simultaneously. Additionally, multimodal models have the benefit of retaining more information when working with different types of data, as they can directly process each form. For instance, they understand speech in its natural state, capturing nuances such as tone and speed without needing transcription. The multimodal approach is poised to become increasingly more important due to its versatility and power (Gozalo-Brizuela & Garrido-Merchán, 2023).

Others:

Generative AI has many more use-cases than those already mentioned. Applications in tasks such as decoding brain activity into pictures and text (Sun et al., 2023), modeling protein structures, designing drugs, and training data generation for self-improvement are only some of the diverse possibilities (Y. Cao et al., 2023; Gozalo-Brizuela & Garrido-Merchán, 2023; Nalini et al., 2023).

1.6 Challenges and Limitations of Generative AI

While offering transformative potential, Generative AI also faces challenges and limitations. These include data biases, harmful content creation, opaque decision-making processes and other concerns such as authenticity and misuse (Weidinger et al., 2023). Additionally, the technology's limited adaptability to new, unforeseen scenarios presents a significant hurdle.

The recognition of these challenges naturally leads to a deeper inquiry into the implications of generative AI. Thus, the following chapters critically analyze these dimensions.

2 Distinguishability of AI Content

One of the limitations generative AI faces is its ability to produce outputs indistinguishable from human-made content. This emerging capability challenges human capacity to discern between what is created by humans and what is generated by machines, particularly raising questions about the authenticity of media. The specific implications of this issue is examined in subsequent chapters.

The Turing Test, introduced by the British mathematician Alan Turing in 1950, was originally conceived as a method to determine the presence of intelligence in a machine (S. J. Russell et al., 2022). The original test is passed if a human evaluator, engaging in conversation, cannot reliably tell whether they are interacting with a human or a machine. Over time, however, the consensus has shifted in the AI community. It is now widely acknowledged that passing the Turing Test does not necessarily prove the existence of genuine intelligence, as the test primarily measures a machine's ability to mimic human-like responses (Dodig-Crnkovic, 2023; Pischedda et al., 2023). This approach aligns with the "acting humanly" category of AI, whereas the current focus has shifted towards the "acting rationally" category, which is more broadly accepted (S. J. Russell et al., 2022).

Despite this shift towards the “acting rationally” category, the general approach of the Turing Test remains relevant, particularly in the context of advanced models like ChatGPT, which demonstrate a sophisticated ability to imitate human conversation realistically in text (Eke, 2023). This raises significant questions about the discernibility of AI-generated content. While the original Turing Test focused on verbal interactions, the scope of AI has expanded to include various other modalities. This chapter explores the extent to which AI-generated content in different modalities – including text, images, video, and sound – can currently still be distinguished from human-created content.

It is also crucial to consider the level of autonomy present in current AI models. This involves examining which modalities can be generated in real-time and interactively, and which require human review to ensure they are indistinguishable from human-created content.

Text:

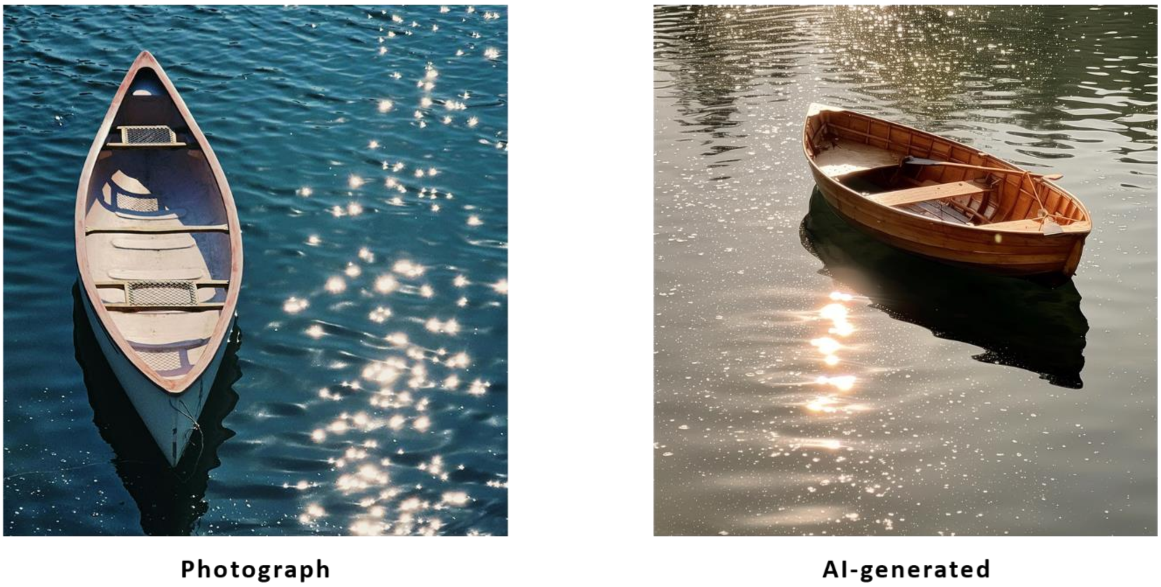
As previously mentioned, it is already possible to generate text that is indistinguishable from human-written text. This is achievable in real-time and interactively, thanks to the capabilities of language models like ChatGPT. These models can respond and adapt to inputs on-the-fly (Dodig-Crnkovic, 2023).

Image:

AI-generated images can already keep up with realistic human-made photos. However, the quality is usually dependent on approach and human moderation to ensure the generation of realistic depictions (Göring et al., 2023). Furthermore, high-quality editing of existing photos is also already possible already (Yildirim et al., 2023). In the case of stylized pictures, less moderation is needed, assuming the subject matter is not overly complex or includes text elements, which can contain generation artifacts.

This is due to the fact realism demands a greater number of factors to be indiscernible compared to less constrained formats. The time required for generation is continuously decreasing (Y. Cao et al., 2023), often surpassing human capabilities in terms of speed. New developments are also allowing to create high quality images in practically real time (Pernias et al., 2023; Rampas et al., 2023; Trevithick et al., 2023).

Figure 1 Comparison photograph (Pexels) and realistic generated picture (Midjourney)



<u>Sound:</u>

In the field of sound, realistic quality speech synthesis has seen significant advancements as well. The generation can be done fast and interactively (Mahajan et al., 2021; Wang et al., 2023). *”It is now possible to generate synthetic speech of such a high quality that is indistinguishable from real voices by human listeners”* (Cuccovillo et al., 2022, p.1). Notably, the cloning of existing voices is possible as well (Arik et al., 2018). In music generation, AI also already has the capacity to produce compositions, especially in less structured or more abstract musical forms. However, achieving the level of quality found in high-end human compositions is remains unachieved due to several limitations (Zhu et al., 2023).

<u>Video:</u>

Currently, no publicly available AI model can generate realistic videos solely from text input, but OpenAI’s recently announced model Sora is able to do so under human supervision, as well as animating static images and more (OpenAI Reseach, 2024). Furthermore, there are AI-based video editing techniques that yield realistic results. A prominent example is the use of deepfakes, where AI algorithms can convincingly swap faces in videos (Chadha et al., 2021; Westerlund, 2019). This can already be achieved in real-time (Fan et al., 2022). Realistic video dubbing, where AI synchronizes lip movements with altered audio, is another advancing area (Xie et al., 2021). These processes usually require human oversight and are not fully automated. Regarding the generation 3D video without moderation, achieving realism remains a challenge, yet significant progress is being made (Kerbl et al., 2023). On the other hand, the creation of stylized video content is progressing rapidly.

In modalities not aimed at replicating reality, such as coding or other applications, the issue of distinguishability between human and AI-generated content becomes less critical. Instead, the emphasis in these domains is on the creativity and quality of the output, rather than on the ability to mimic reality. However, a new level of indistinguishability may emerge if AI models begin to surpass human performance in tasks traditionally considered human-specific, like coding. While AI models are becoming increasingly capable, they have not yet achieved complete autonomy. Their ability to operate independently and make decisions without human input is advancing, but there are still notable limitations in their autonomous capabilities (Mezrich, 2022).

3 Ethical and Societal Implications

The increasing difficulty in distinguishing between human and machine-generated content, as well as other aspects of generative AI raise significant questions. This chapter delves into the concerns brought forth by the emergence of this technology, exploring some of the ethical and societal implications.

3.1 Content Provenance and Trust in Information

As technology continues to advance, the ease of creating realistic content across various modalities is rapidly increasing. With AI requiring less labor than traditional methods and its growing autonomy, we are approaching a future where AI-generated content could dominate digital realms like the internet (Bergman, 2022). Some even estimate as much as 90 percent of media on the internet could be synthetic by 2026 (Schick, 2020).

In an environment where a significant portion of content may not be authentically human, establishing reliable solutions for content provenance is crucial, helping users to distinguish between synthetic and human media. This level of transparency is essential not only as an ethical practice but also for maintaining public trust in critical information sources. As the origin of content becomes increasingly ambiguous, uncertainty and mistrust might emerge, complicating the public's ability to discern fact from fiction (The Act | The Artificial Intelligence Act, 2021; Vaccari & Chadwick, 2020). Early research suggests that synthetic content is often perceived as equally or even more trustworthy than human-generated content (Huschens et al., 2023; Kreps et al., 2022). The risk here is the potential for being misled by media that appears real, leading to misinformation and increased costs and efforts in acquiring accurate information.

Furthermore, if content perceived as trustworthy is inauthentic and potentially misleading, this could further erode trust and increase uncertainty in information, particularly on social media platforms. There is already evidence suggesting that low political trust correlates with the use of digital media (Lorenz-Spreen et al., 2023). This may partly be due to inadequate moderation of content for factuality, leaving users without clear evidence of correctness. Traditional methods of fact-checking may struggle to keep pace with the sheer volume of synthetic content spread. This widespread uncertainty may cause any digital evidence like videos to become illegitimate, which could be exploited, for instance, by political figures who might deny accusations by casting doubt on the accuracy of claims, as the truthfulness of such claims becomes increasingly difficult to prove (Vaccari & Chadwick, 2020)

> *Experts fear this may lead to a situation where citizens no longer have a shared reality, or could create societal confusion about which information sources are reliable; a situation sometimes referred to as 'information apocalypse' or 'reality apathy'."*(European Union Agency for Law Enforcement Cooperation., 2022)

The implications of broadly spread AI-generated content are profound. The need for robust mechanisms to verify content authenticity and origin becomes visible. This challenge underscores the importance of developing and implementing effective strategies for content verification and transparency in the age of generative AI to ensure that the public can trust the information they encounter.

3.2 Transparency Issues and Responsibility Assignment

Assigning responsibility for AI systems and their potentially faulty behavior presents unique challenges compared to traditional digital technologies. One primary issue is the complexity of state-of-the-art AI models, which often employ deep learning techniques. These models mostly function as 'black boxes,' where the internal decision-making processes are opaque. While the input and output are observable, the mechanisms within remain hidden (Rai, 2020). This lack of transparency complicates the assignment of responsibility and makes it difficult to pinpoint the exact cause of an issue or error (The Act | The Artificial Intelligence Act, 2021).

Another factor complicating the assignment of responsibility is the autonomy of AI systems. These systems can make decisions increasingly independently, without human intervention. Additionally, they can evolve over time, leading to unpredictable outcomes. This autonomy further obscures the attribution of responsibility, as most legislations do not cover non-human actors and the systems' actions may not reflect the intentions or control of their creators (Orr & Davis, 2020; Turner, 2019).

AI developers themselves acknowledge the severe limitations they face in controlling AI systems post-deployment and the corresponding accountability challenges:

> *Uncovering algorithmic authorship and ensuring technical transparency is rarely sufficient nor feasible to elucidate attributions of ethical accountability. Blame does not rest easily with designers, users, hardware, or code, but rather, somewhere in the spaces between.* (Shank et al., 2019, as cited in Orr & Davis, 2020, p. 4) .

This statement highlights the complexity of assigning blame or responsibility in the context of AI. It remains unclear who should be held accountable for ensuring the ethical character and embedding standards of fairness, particularly when developers themselves struggle in the absence of uniform norms.

AI systems involve multiple actors throughout their lifecycle, including conception, design, implementation, and use. Responsibility for AI systems' outcomes is to be distributed among these various actors, making it challenging to identify a single entity, stakeholder, or individual as solely responsible (Orr & Davis, 2020).

3.3 Consent and Intellectual Property

In modern machine learning and particularly generative AI, data is a critical component, categorized into input (used for training the models) and output (generated by the models). As current models require vast amounts of data, it has become increasingly common to use training datasets that include publicly available data, often originating from the internet (S. Huang & Siddarth, 2023). However, the use of public data raises ethical concerns, particularly when such data contains copyrighted or private information. Research has shown that models like ChatGPT can leak such training data (Chang et al., 2023; Nasr et al., 2023).

As previously noted, there is a lack of transparency in complex models, which makes it extremely difficult to trace back which input data points were used to create the final output (Rai, 2020). This becomes problematic, especially when the input data is used without explicit consent.

For instance, when AI models are trained on art or music created by specific individuals, the resulting AI-generated content might replicate or derive from these original works. This situation raises complex questions about intellectual property rights and copyright infringement as well as authorship and ownership, particularly when considering who holds the rights to AI-generated content. The current legal frameworks are not satisfactory to regulate these cases. Therefore, new regulations are necessary (Watiktinnakorn et al., 2023).

Generative AI poses challenges not only to individual property rights but also to collective properties, such as the 'digital commons.' This term refers to the online resources that society collectively owns, benefits from and contributes to, including wikis, internet archive snapshots, as well as Creative Commons materials. With machine-generated content likely to become a significant portion of these digital commons, there is a potential risk of decreased content quality. (S. Huang & Siddarth, 2023).

> *Such issues may greatly degrade the quality of the information commons and require some level of restructuring of the internet e.g. intensifying and requiring new solutions to the problem of how we detect, filter and rank machine- vs human-generated content.* (S. Huang & Siddarth, 2023, p. 6).

Another aspect is the potential for misuse, as it is substantial, specifically regarding already legally restricted actions. A good example is the creation of deepfakes or other forms of digital imitation of individuals as e.g. in chatbots (Lee et al., 2023). While this technology has positive applications, it is frequently exploited for harmful purposes, such as creating nonconsensual pornography or spreading misinformation (Chadha et al., 2021; European Union Agency for Law Enforcement Cooperation., 2022; Westerlund, 2019).

3.4 Biases

Bias and discrimination are common concerns in AI systems, especially regarding more complex systems. These biases often originate from the data used to train these models. Since there are many biases in reality, the training data frequently reflects already existing biases. These biases then reoccur in the outputs of the systems (Li et al., 2021).

There have been documented instances of generative AI models like modern LLMs or Image-generation models producing biased content containing bias towards gender (Bolukbasi et al., 2016; Lucy & Bamman, 2021), race (Hosseini, 2023), disability (Hutchinson et al., 2020), religion (Abid et al., 2021), and societal status (Bender et al., 2021). These biases occur in various ways, from amplifying stereotypes to providing unequal or unfair treatment to certain groups (Bianchi et al., 2023).

As generative AI is typically trained on data from the internet, wealthier communities and countries are overrepresented. This distortion in data representation leads to AI systems that are less effective or relevant for underrepresented populations. The result is a digital divide where AI technologies are more attuned to the needs and contexts of certain groups over others (Bender et al., 2021).

The elimination of these discriminative effects may be virtually impossible. This is due to the fact that these normative evaluations of acceptance thresholds are not objective and have to be made by humans and vary from context to context (Bianchi et al., 2023; Weidinger et al., 2023).

3.5 Education

The potential of generative AI to improve learning and education capabilities is immense, even so far as to a complete revolution in academia. AI can assist educators in creating more personalized and inclusive learning experiences. This could include simple applications like fast and flexible translation services to more advanced functionalities such as the personalization of educational materials. As already discussed, AI has the potential to adaptively tailor content, therefore optimize tasks and content to individual students, enhancing understanding, providing deeper insights, and offering additional guidance (Grassini, 2023; Rudolph et al., 2023).

One promising application is the use of AI for grading students and giving them feedback. While this is currently mostly viable for relatively uniform tests, AI's capability to evaluate complex subjects is rapidly advancing. This technology could offer more thorough and unbiased feedback compared to traditional methods, reducing personal evaluator biases and labor time limitations (Grassini, 2023).

Looking into the more distant future, digital AI educators could make education more accessible, particularly for students with disabilities or those in remote areas. This democratization of education could bridge significant gaps in the current system (Zdravkova et al., 2022).

However, with the integration of AI in education, risks emerge, as well. One critical concern is the potential undermining of academic integrity. The ease with which students can outsource assignments to AI tools, from writing to problem-solving, raises several questions. Some traditional assessments, like essays, may become obsolete unless new, AI-appropriate approaches are developed (Eke, 2023; Rudolph et al., 2023). The over-reliance on AI tools for learning tasks could lead to a decrease in learning and understanding, similar to effects observed in research with frequent use of modern GPS navigation systems. These systems, while providing step-by-step directions, negatively influence spatial memory, potentially due to reduced involvement in the decision-making process (Dahmani & Bohbot, 2020).

In terms of authenticity regulation, there are currently no reliable methods to detect AI-generated content, although potential solutions are being explored. The previously discussed issues of intellectual property and plagiarism are particularly prominent in academia. Thus the community has to reevaluate what constitutes knowledge and how it can be acquired in a time where seemingly clear borders blur (Eke, 2023).

3.6 Power Dynamics and Monopolies

Not long ago, AI research was mostly a collaboration endeavor between academia and industry, with both sectors contributing significantly to the field. However, since the advent of the deep learning era around 2012, the industry has increasingly taken the lead. This shift is largely due to the huge data and computational power requirements of deep learning systems. Initially, only top-tier universities could compete with industry giants, but recently, the development of the largest and most advanced models is predominantly driven by private companies (Ahmed et al., 2023; Ahmed & Wahed, 2020).

The recent shift in attention towards AI, particularly with the deployment of models like ChatGPT, has ignited a competitive race among major tech companies for dominance in this critical field (LaGrandeur, 2023). Companies that were already resource-rich have a distinct advantage, as they already possess the necessary resources (Ahmed et al., 2023). This disparity risks creating AI and data monopolies, as these companies solidify their positions in the market. Particularly, a breakthrough in AI could leverage the influence and power of the developing entity and lead to a monopoly (Liu, 2022).

3.7 Economics and Skills

"Generative AI is likely to impact just about every industry" (Taulli, 2023, p.188). This drastic statement aligns with the rapid advancements and foreseen potential of generative AI technology. As AI systems become increasingly sophisticated, their influence across various sectors is accelerating (Selenko et al., 2022). It is estimated that the integration of generative AI into business and society could increase the global GDP by seven percent in a ten-year period. While some industries may experience less impact than others, projections do suggest that up to a quarter of all jobs in the US and Europe could be

automated by AI. Globally, this could sum up to approximately 300 million jobs being replaced (Goldman Sachs Research, 2023).

Historically, technological advancements have not only led to worker displacement, but also to the creation of new job opportunities. However, the trend in recent years, particularly since the rise of IT automation, has shown a net decrease in labor demand. While AI will affect all industries, currently it is more likely to augment existing roles rather than completely replacing them. This partial automation shapes a complementary relationship between AI and human workers (Jan Hatzius et al., 2023).

In the near future, workforces with ordinary digital skills are particularly at risk, as digital technologies rapidly acquire and surpass these skills (West, 2018). While automating repetitive and mundane tasks could potentially increase overall job satisfaction, there is also a concern that a higher degree of automation might accelerate the pace of work. This acceleration could, in turn, lead to decreased job satisfaction, as workers may face increased pressure and demands in the automated work environment (Bommasani et al., 2022).

As generative AI models evolve to perform increasingly complex tasks and potentially become a general-purpose-technology/general AI, the demand for traditional human labor and skills may diminish. Advanced AI, not merely limited to cognitive tasks anymore, could enable a small number of experts to perform tasks more efficiently, effectively amplifying their capabilities (Cramarenco et al., 2023; Poba-Nzaou et al., 2021).

Regarding a more distant future where machines do most of the work, and unemployment becomes more common, fundamental reevaluation of our resource distribution systems will be necessary. With adequate measures, a more satisfactory life with additional leisure could be the result (West, 2018).

3.8 Disinformation and Political Manipulation

As established earlier, AI has the capability to create content indistinguishable from content produced by humans. While not all AI-generated media are inherently misleading or harmful, the decreasing resources required for media generation allow the fast and cost-effective creation of disinformation material (S. Huang & Siddarth, 2023). Resources required to generate large volumes of authentic-seeming content, without necessarily compromising quality, are minimized. This ease of production enables both malicious individuals and larger organizations to spread disinformation more effectively and efficiently by scaling up and optimizing their disinformation campaigns (Goldstein et al., 2023).

Regarding political manipulation, the 2016 incident involving Cambridge Analytica, a British political consulting firm, highlighted the potential and intent to manipulate public opinion for political purposes using digital media, particularly in the context of elections. By spreading tailored, misleading advertisements on Facebook, Cambridge Analytica influenced users' voting preferences to interfere with the 2016 US election (Hinds et al., 2020). While such tactics were effective in 2016, the increasing capabilities of generative AI has the potential to magnify such actions significantly.

Research indicates that humans are often more easily deceived by machine-generated content, compared to content created by humans. This susceptibility increases the potential for harm caused by this technology. Furthermore, the effectiveness of disinformation is frequently linked to its emotional impact. Generative AI, with its ability to produce personalized and therefore potentially emotionally resonant content, can be particularly effective in this regard (Weidinger et al., 2023).

High-quality synthetic content, such as deepfakes or texts generated by advanced LLMs, can be particularly persuasive. They appeal to both rational and emotional aspects of human cognition, and visual content, which is often perceived as a direct reflection of reality, can be especially misleading. Consequently, these sophisticated forms of synthetic content can more easily lead to false beliefs about reality (Vaccari & Chadwick, 2020; Weidinger et al., 2023).

Interestingly, the effectiveness of such campaigns does not always hinge on the quality of the content. In scenarios where the goal is to sow confusion rather than convince, the sheer volume of posts can be more impactful than the quality of each piece. A flood of diverse, contradictory, and nonsensical information, even if partially spread by malicious actors, can create a state of collective uncertainty and distrust. It becomes challenging to discern the truth, and verifying the authenticity of information can be a difficult task (Goldstein et al., 2023; S. Huang & Siddarth, 2023; Vaccari & Chadwick, 2020).

These capabilities are not limited to entities with political motives. Whilst propaganda for hire or governmental manipulation are some of the more concerning applications, personalized synthetic content has many more potential uses (Goldstein et al., 2023).

3.9 Personalized Synthetic Content and Psychological Impact

Particularly regarding personalized content, AI's capabilities far surpass what is feasible with human labor alone. Tailoring content to a wide array of individual preferences and interests is a task that would be impractical, if not impossible, for humans to perform at scale.

This personalized content, generated by AI, can include various modalities and forms. It can manifest as targeted advertisements, customized emails, or even personalized chatbots that interact with users on an individual basis. For instance, AI-driven chatbots can be programmed to emulate a diverse range of personas, making them especially relevant in social media contexts. (Bommasani et al., 2022; Goldstein et al., 2023).

Thus, the malicious use of such personalized AI-generated media is not limited to political motives, as discussed in the context of the last chapter. The applications are varied and could include such as persuasive advertising, phishing attacks and forms of social engineering, among other uses (Goldstein et al., 2023). For example, the emerging marketing phenomenon of virtual influencers has already been shown to be a powerful tool to alter users' preferences. They are often perceived as attractive, more trustworthy, and more credible (Gerlich, 2023). Naturally, the content may be generated without any malicious intent at all. Therefore, these individualized media can also be useful for users to receive content, which is adapted to each person (Goldstein et al., 2023).

The frequent interaction between humans and generative AI models, particularly chatbots, is set to profoundly influence social and psychological dynamics. This impact, while seemingly speculative, becomes evident upon examining existing human-computer interaction phenomena. Historically, even basic interactions with computers have inherited social aspects, not necessarily tied to any belief in the social nature of these systems. Social responses to computers are common and occur naturally, often without conscious reflection (Nass et al., 1994).

This tendency is amplified when AI systems imitate human behavior more closely. *"This became known as the "ELIZA effect," the propensity for humans to ascribe understanding and intelligence to computer systems"* (Berry, 2023, p. 11). Originating from one of the earliest chatbot programs, this effect illustrates how people project human attributes onto computers. Such projections can lead to parasocial relationships with AI entities like chatbots, where individuals feel engaged in a personal, reciprocal interaction with a media persona (Gerlich, 2023).

The implications of these relationships are diverse. Chatbots, while potentially consulted due to loneliness, can themselves consequently contribute to greater social isolation. The constant availability and support of AI can surpass human interaction in appeal (Diederich, 2021).

> *If emotional support from a software program is accepted, then no human friend or partner can compete with the endurance and consistency of support provided by a computer. This is an unquestioning and enduring support, that obviously is not mutual, and has an addictive potential.* (Diederich, 2021, p. 38).

Another emerging possibility is creating digital clones of existing persons or shape the personality of the digital entity to individual needs, which could further intensify the addictive potential and raises concerns about personal rights (Lee et al., 2023). Such reliance on AI for emotional support can create dependency or reduce stress resistance, when emotional assistance is constantly available. Users may have to learn how to live independently (Diederich, 2021).

Moreover, interactions with AI can create reinforcing feedback loops, similar to echo chambers on social media (Del Vicario et al., 2016). If a Chatbot is designed to prolong user engagement, as e.g. OpenAI is planning to do with custom variants of ChatGPT as a basis of their builder revenue program (OpenAI, 2024), it might consistently affirm and validate user opinions. This could lead to the solidification or even radicalization of beliefs, particularly in individuals who are socially isolated.

But conversational AI technologies do also offer a unique potential to democratize access to therapeutic guidance, bypassing limitations such as the scarcity of therapists and facilitating open communication in scenarios where individuals might hesitate to engage with human counselors (Gupta et al., 2022).

4. Regulation and Control

Many of these examined aspects clearly reveal the potential consequences that emerge with the use of generative AI. Clearly, there are many positive implications as well, but the mitigation possibilities of the occurring risks and threats must be adequately addressed. Therefore, this chapter dives into the available or upcoming options to regulate and control generative AI.

4.1 Current approaches to AI regulation

AI technology has rapidly integrated into the personal lives of many humans. Though entailing a variety of uncertainties and risks, a comprehensive regulatory framework for AI has not yet been established anywhere around the globe. Before investigating potential solutions for specific risks, it is crucial to examine the current regulatory approaches of major global players. These legislative efforts are exclusively led by governmental organizations, as they are the key actors in standardizing the addressment of these concerns, at least up to this point.

Several initiatives have been launched by leading entities such as the United States, China, and the European Union, each with different focal points. The European Union's approach with the AI Act is the most relevant in the context of mitigating potential ethical or societal risks. It focuses on personal data protection and regulating high-risk AI applications (Luckett, 2023).

The European Commission aims to ensure 'trustworthy AI' by categorizing AI systems based on their potential risk: minimal, high, and unacceptable. Systems with minimal risk, such as AI-enabled recommender systems or spam filters, will not be subject to regulation. High-risk systems, however, will be *"required to comply with strict requirements, including risk-mitigation systems, high quality of data sets, logging of activity, detailed documentation, clear user information, human oversight, and a high level of robustness, accuracy and cybersecurity"* (European Commission, 2023, p. 1).

An example of a high-risk system could be those systems which are operating within critical infrastructure. The third category, unacceptable risk, includes AI systems that pose clear threats to fundamental human rights and will be banned. Whilst violation of any rule will be penalized, specifically the use of such banned tools will be subject to heavy fines. This includes systems designed to manipulate human behavior or subvert free will, such as those used for intentional election manipulation, as previously discussed.

The Commission also addresses transparency risks, introducing the rule, that users should be aware that they are interacting with a machine. Deepfakes and other AI generated content will have to be labelled as such, and users need to be informed when biometric categorization or emotion recognition systems are being used. *"In addition, providers will have to design systems in a way that synthetic audio, video, text and images content is marked in a machine-readable format, and detectable as artificially generated or manipulated"* (European Commission, 2023, p. 1). This approach is further examined in a latter subchapter. Moreover, the Commission plans to impose stricter obligations on general-purpose models that could pose systemic risks and is establishing a new European AI Office to enforce binding AI rules and ensure coordination (European Commission, 2023).

China's regulatory focus is largely on achieving AI supremacy. While the United States shares this goal, it also acknowledges the need to address ethical risks. The United States are promoting the development of 'trustworthy AI' characterized by safety, reliability, and transparency. Their guidelines emphasize the importance of transparency, fairness, and accuracy in AI systems, but they do not yet impose as strict rules as the European Union (Luckett, 2023).

4.2 Maintain Distinguishability of AI Content

As discussed in a previous chapter, ensuring that synthetic content is discernible from human-generated content is a critical concern, as severe implications emerge with non-distinguishability. There are multiple potential approaches to mitigate these risks and threats, which are now examined.

Preventive Measures in AI Development:

Starting earlier at the creation phase, AI developers and companies could aim to build models that inherently produce more detectable content. This strategy requires a high level of coordination and is not yet proven to be technically feasible. A significant limitation is that only models intentionally altered to create detectable output will do so. This limitation is reoccurring across multiple strategies, as coordination is a key element in shaping the media landscape of the future. It is unrealistic to expect universal participation in such programs, and malicious actors may gravitate towards unaltered models (Goldstein et al., 2023).

"Radioactive" Training Data:

Exploring further, one technique in development, happening before content generation, involves using "radioactive" training data, data with certain attributes which can later help identify the origin of the outputs. As little as one percent of the training data might be sufficient for easier detectability. However, with rapidly increasing model sizes, this approach might face scalability challenges. There is also technical uncertainty related to this technique. Altering the outputs after generation could diminish detection capabilities, making the approach potentially easy to circumvent. The method of introducing this data into the training sets varies. It could be implemented directly by the training entities or indirectly planted through other means, such as releasing it onto the internet for models that use data gathered through web crawling. This latter approach could have a more substantial impact, affecting a wider range of models, but it comes with ethical concerns and uncertainties. The intentional publication of huge amounts of data, the content of which is unknown, is not to be done without delicate consideration. In the end, it is also supposed to be a choice made voluntarily by the training entities, and the legality of indirectly forcing them to participate is another critical prospect (Goldstein et al., 2023; Sablayrolles et al., 2020).

Watermarking Techniques:

Another approach is watermarking, which differs from radioactive data as it occurs in the end of the generation process and is not related to the training dataset. Google DeepMind's SynthID is a recent example, integrated into their image generation model Imagen. In the context of images, this technique embeds imperceptible identification markers into the pixels of pictures generated with their model. Even if the images are altered and metadata is removed, the identifier can remain detectable, as embedded in the individual pixels, though not guaranteed. The reidentification works via a confidence value assessing the likelihood that the content is generated. However, like the radioactive data approach, this technique's effectiveness relies on AI companies' participation (Gowal & Kohli, 2023).

Platform and Developer Collaboration:

Platforms and AI developers could collaborate, coordinating efforts to identify AI content. They might enforce rules requiring the disclosure of content origins, which, although evadable, could reduce the overall volume of non-disclosed synthetic content and increase awareness. Automatic detection, like YouTube's Content ID framework, might face feasibility issues, especially if platforms use encryption. The effectiveness of this approach is once again limited to participating entities (Goldstein et al., 2023).

Proof of Personhood:

Requiring proof of personhood for posting content is another strategy. Verification methods contain a wide range of sophistication. From providing basic personal information like full name and phone number to more sophisticated systems like CAPTCHAs and ID verification. The currently most reliable method involves video-proof of people and official identification documents like national IDs and passports, but this approach faces upcoming challenges with the emergence of convincing live-deepfakes and frequent identity fraud (Goldstein et al., 2023). Additionally, reliance on official documents could exclude around 850 million individuals without these documents, mostly citizens of developing countries (World Bank, 2023). Though, the verification of personhood still would be no verification of the human origin of content. While no method is foolproof, more stringent verification can make radically increase the effort required for automated spreading of synthetic content. However, this approach raises privacy concerns and could impact free speech (Goldstein et al., 2023). And once more, there is a reliance on the platforms enforcing such terms in the first place. As this might not be the most economically rational decision, platforms could choose to not participate.

AI-based Detection Tools:

Utilizing AI tools for detection of AI-generated content is an intuitive countermeasure. These tools are becoming more sophisticated in determining content origins, but there are strategies to make content less detectable, as well. Some techniques are able to detect AI content, specifically deepfakes with reasonable reliability, by examining creation artifacts (H. Huang et al., 2022). There are already tools on the market which claim to be able to detect whether certain texts emerged from LLMs, as well. Even though often quite unreliable, Plagiarism checkers, such as Turnitin, commonly used in academia, already include such features, as well. But they are improving, and this could lead to a continuous cycle where generation tools stay a step ahead of detection tools. Retrospective detection might be possible, but if the content is harmful, the damage might already be done (Elkhatat et al., 2023; Weber-Wulff et al., 2023).

Consumer-Focused AI Tools:

Consumer-focused AI tools could focus on another matter at an earlier point of time. *"While detection methods aim to detect whether content is synthetic, consumer-focused tools instead try to equip consumers to make better decisions when evaluating the content they encounter."* (Goldstein et al., 2023, p. 61). These tools can provide information about media and hints about potential origins, enabling users to critically assess content. However, these tools may inherit biases similar to their generative counterparts (Goldstein et al., 2023).

Digital Provenance Standards and Media Literacy:

Adopting digital provenance standards, as proposed with the European Union AI Act, and promoting media literacy campaigns can raise awareness and aid in detecting non-genuine content. While limited in effectiveness, these strategies contribute to a bigger picture of issues surrounding synthetic content (Goldstein et al., 2023; The Act | The Artificial Intelligence Act, 2021).

In summary, while there are many strategies to enhance the detectability of AI content, a comprehensive solution to this issue remains a challenge. The effectiveness of these approaches will likely evolve over time, as both AI technology and regulatory measures continue to develop.

4.3 Transparency and Trust

The ability to understand the decision-making process of AI systems is crucial for many prospects, particularly for fostering human trust in these technologies. As many modern deep-learning-based AI systems process humanly unmanageable amounts of data, they are therefore typically not interpretable by humans. This inherent "black-box" design has incentivized extensive research into making AI's internal processes more transparent to external observers (Haresamudram et al., 2023). The urgency of this issue was further highlighted by the European Union's legislation, which requires all automated individual decision-making to be explainable when significantly affecting users (Goodman & Flaxman, 2017). These concerns led to the emergence of a new field: Explainable Artificial Intelligence, or short XAI, where AI algorithms consider human comprehension of their outputs (Haresamudram et al., 2023).

Current approaches for the black-box introspection include model explanation, outcome explanation, and model inspection.

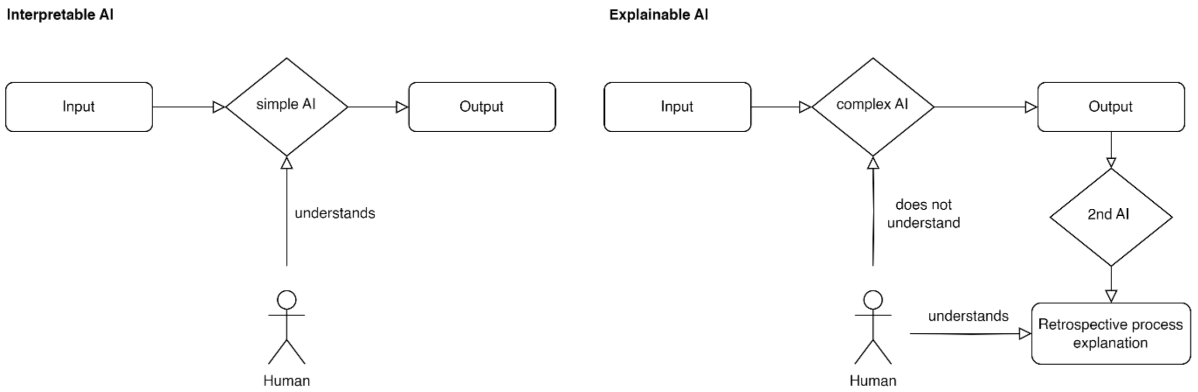
> *"Model explanation consists in building an interpretable model to explain the whole logic of the black box while outcome explanation only cares about providing a local explanation of a specific decision. Finally, model inspection consists of all the techniques that can help to understand (e.g., through visualizations or quantitative arguments) the influence of the attributes of the input on the black-box decision"* (Aivodji et al., 2019, p. 1).

All three these explanation types, generated by secondary AI algorithms, aim to decipher the decision-making process of the primary AI system retrospectively. While these explanations can be useful for understanding, their reliability is sometimes questionable. They often offer plausible and believable narratives, focusing more on the interaction between the AI and the user than on the algorithm itself (Haresamudram et al., 2023).

Indeed, it has been demonstrated that it is *"possible to systematically rationalize decisions taken by an unfair black-box model using the model explanation as well as the outcome explanation approaches"* (Aivodji et al., 2019, p. 1). This phenomenon, labeled "fairwashing," involves creating a false perception of fairness in AI decision-making, even when the underlying decisions are biased or unfair (Aivodji et al., 2019). Such possibilities could further decrease the trust in synthetic content, as inauthentic transparency of the internal processes undermine the potential transparency of any model. The recognition of such risks underscores the need for ongoing research into the transparency of modern AI systems.

An alternative to XAI is the development of Interpretable AI, which utilizes less complex human-comprehensible algorithms instead of opaque black-box models. In high-stakes applications, where the implications of AI decisions are significant, the use of these "transparent-box" systems is often recommended (Rudin, 2019). But as their size is limited, these systems do have limitations in capabilities.

Figure 2 Illustration of Interpretable AI and Explainable AI



4.4 Responsibility

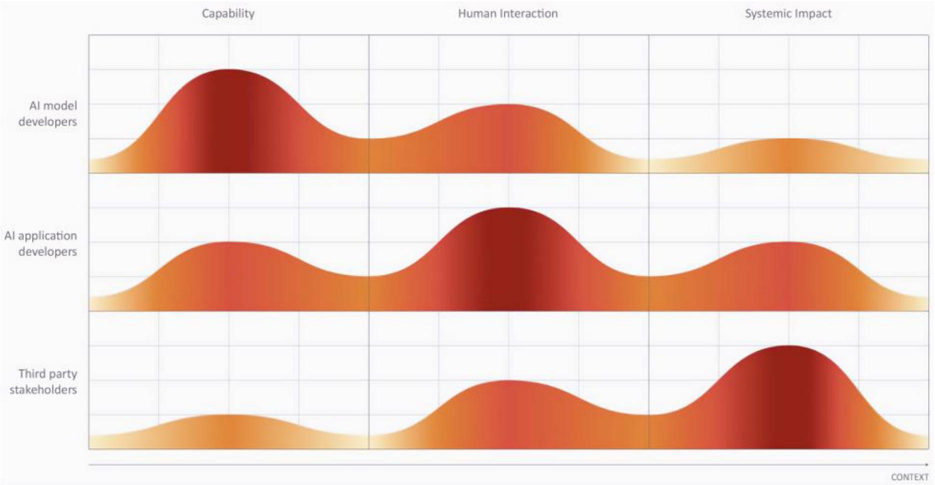
As the outputs produced by models can be unpredictable and the issue of lacking transparency is prevalent, distributing the responsibility for AI systems must start before public deployment.

Firstly, the primary responsibility lies with the AI developers. They must rigorously test and understand the capabilities and potential risks of their models (Weidinger et al., 2023). It is their duty to minimize the creation of potentially harmful material. However, given the inherent unpredictability and complexity of AI systems, completely preventing harmful content or behavior is a challenging task (Tate, 2023).

Secondly, application developers and specific public authorities hold accountability for the functionalities and features of AI applications. They must ensure that these applications follow ethical standards and regulatory requirements regarding different user groups and use-cases.

Moreover, broader public stakeholders, such as governments and civic interest groups, play a crucial role in assessing and responding to the societal implications of AI technologies. Their input is essential in shaping the ethical landscape in which AI operates (Weidinger et al., 2023).

Figure 3 Graph relative distribution of responsibility across multiple layers and contexts (Weidinger et al., 2023)



Finally, users of AI systems also own a degree of responsibility. They should use these tools in accordance with the terms of service and for intended purposes. While AI products must undergo thorough testing before public deployment, users must also act responsibly, otherwise they can be held accountable for the results (Orr & Davis, 2020).

In summary, the responsibility for safe and ethical AI systems is a currently shared burden, requiring collaboration across various sectors and stakeholders.

Looking into the future, advanced AI systems with a reasonable degree of autonomy in their decision-making might be assigned responsibility themselves, independently of their sentience (Turner, 2019).

4.5 Intellectual Property and Copyright

Legislative Regulation:

AI-generated media currently exist in a precarious legal gray area regarding copyright (Watiktinnakorn et al., 2023). *"Generally speaking, legal systems do not provide for copyright-protected works being created by non-humans"* (Turner, 2019, p. 123). Only few legislations specify copyright for cases without a human creator. The UK Copyright, Designs and Patents Act 1988, for example, states: *"In the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken"* (Copyright, Designs and Patents Act 1988, 1988, p. 1). However, the definition of the person who made the arrangements for AI-generated media remains vague. It could be the person who built the system, trained it, or provided specific inputs. Looking ahead, increasingly capable and autonomous AI systems might one day be attributed rights themselves (Turner, 2019).

Adjusting legislation to accommodate these technological advances is pressing, especially as AI systems significantly contribute to modernizing many processes and improving possible results, often in creative collaboration with humans. Two opposing perspectives have emerged regarding future copyright regulation. One side advocate for extending copyright protection to AI-generated works, arguing that these creations exhibit a level of creativity and innovation and are on par with human creations, thus warranting legal safeguarding. They demand that at least works resulting from joint human-AI efforts should be included under copyright law.

Conversely, others argue that AI-generated works lack genuine creativity due to minimal human input and ingenuity, disqualifying them from copyright protection (Watiktinnakorn et al., 2023).

The US Copyright Office has recently denied copyright protectability of AI-generated images, providing some precedent and declaring synthetic media as public domain (Brittain, 2023). However, the blurred lines between collaborative human-AI works complicate this stance. The Copyright Office has initiated a study to investigate these complex circumstances further, indicating that these cases do not yet have long-ranging implications (Library of Congress, Copyright Office, 2023).

Technical Approaches to Protecting Intellectual Property:

Several technical strategies are being explored to protect intellectual property in the context of generative AI.

There are techniques conducted on protected images, altering them invisibly, which can disrupt AI-based image translation models, effectively preventing unauthorized modifications under certain conditions.

Watermarking generative models' outputs can indicate not only that the output is synthesized but also identify the specific model used, aiding developers in protecting their property.

Embedding ownership information directly into a model allows models' owners to use a remote trigger to verify if a black-box API is utilizing their models, therefore allowing them to detect non-authorized use of their model (Zhong et al., 2023).

The use of copyrighted data for training new models is still debated and their protection remains a challenge. While the use of such data might seem innocent, as the system does not reflect individual datapoints, the media effectively become part of the output (Gillotte, 2020). This issue is particularly evident with generative models capable of imitating specific styles, affecting many creators as AI reproduces their style, potentially devaluing their original works (Lola Mayor, 2023).

As a response, some models, like DALL-E (OpenAI Reseach, 2021), avoid replicating specific styles or generating individual persons to protect copyright and personal rights. They also allow artists to opt-out their work from the training data, though likely already integrated in current models (Jacob Ridley, 2023). However, this approach is not uniformly applied across all tools; for instance, ChatGPT, another tool from OpenAI, who also own DALL-E, does not have such restrictions and e.g. individual writing styles can be replicated. Some propose the widely adopted financial compensation for imitation of specific styles for artists and contributors (S. Huang & Siddarth, 2023). Adobe Firefly, trained exclusively on licensed material, avoids copyright infringement and even actually compensates contributors (Adobe, 2023), but thus faces competitive disadvantages. Whether the use of copyrighted material is fair is still discussed and different regulations are demanded. If future legislation does not mandate training data to be authorized for processing, mandatory efforts to avoid replicating copyrighted material are demanded (Gillotte, 2020).

This lack of regulation underscores the need for a unified approach. Blockchain technology has been proposed to manage AI-generated content, making it wholly traceable and publicly documented. The Blockchain could assist in ownership verification and provide a reliable environment for AI systems and their outputs to be trained, deployed, and traded legally (Chen et al., 2023).

4.6 Reduce Biases

Reinforcement learning, a practice virtually omnipresent in modern AI models, is not only a powerful tool for optimizing performance but also plays a crucial role in mitigating biases. This approach, whilst many variations exist, frequently involves rewarding the model based on the desirability of its outcomes. It could therefore be compared to the conditioning of animals. It can be implemented automatically, where responses are rated for e.g. clarity, or under human supervision (S. J. Russell et al., 2022). Over time, this process leads to the model's convergence towards more desired, less biased outcomes. This fine-tuning is typically conducted both pre- and post-deployment, with user feedback also being integrated for further refining and continuously optimizing the models. While this approach mostly addresses overtly sensitive topics, it does not inherently eliminate harmful or biased content. This tendency also emerges from the limiting fact that more harmful content is easier to detect (Boxin Wang et al., 2023).

Despite these efforts, no technique currently exists to fully prevent generative AI from producing biased results.

However, additional methods such as the implementation of automatic detection systems are employed to further reduce biased outputs. These systems, integrated into models, are designed to censor harmful and toxic outputs (Hacker et al., 2023). While well-developed for text moderation, these detection methods are less developed for other modalities like image generation (Bianchi et al., 2023). But even in the context of text outputs, not all biased results can be detected, as the bias might only become evident once put into context. As AI technology advances, its capability to understand and interpret complex situations and relationships across different modalities is expected to improve, as well. Therefore, the capabilities of AI to detect toxic content that is less prominent, will increase, too.

Accepting that not all bias can be removed from AI models does not preclude the possibility of further mitigating potential risks. The concept of “Model cards” has been suggested to help disclose how a model performs across variety of aspects, e.g. demographic groups, thereby informing users and researchers about the contexts in which the model is most appropriately used (Kahn, 2022). This approach could prevent the use of models in scenarios where they are likely to exhibit bias.

In high-stakes applications, such as medical diagnostics, there is a growing demand for regulatory bodies to perform a top-down intervention. *“Regulatory bodies and publishing standards must [...] require evidence of objective accuracy”* (Kapur, 2021, p. 460). While achieving and proving absolute accuracy is an extremely challenging task, benchmarks for evaluating the toxicity of model outputs exist. Creating a regulatory frameworks with corresponding thresholds and rules could help to reduce the harm caused by biased AI (Kapur, 2021).

The root of many biases in AI lies in the training data itself. Therefore, optimizing this data is key to creating fairer AI. Approaches include filtering datasets in advance, either manually or automatically. Whilst it is a severely limited approach, it has been shown that certain optimized datasets, containing “textbook quality data,” allow for better performance with the same model size, also in terms of reducing toxicity (Gunasekar et al., 2023).

Similar to the disclosure of known tendencies of models, there is such an course of action for datasets; “Datasheets for datasets” document a dataset's motivation, composition, collection process, and intended uses, potentially increasing transparency and accountability (Kahn, 2022). A frequent limitation of training datasets is the underrepresentation of edge cases or groups. Oversampling these underrepresented data points is a common strategy, but if the original data is biased, it may inadvertently enhance the bias (Kapur, 2021).

Generating new synthetic training data is growing to be common, as well. Though when using the original data for generating new training data, similarly, previously existing biases might be reflected and therefore strengthened. Emerging approaches try tackling this problem by including bias evaluation and adjustment already during the training data generation process, providing already prefiltered, less biased data (Gujar et al., 2022). The exploration of entirely synthetic datasets as a potential solution to bias issues is also underway.

In conclusion, while current strategies have made significant progress in reducing bias and discrimination in AI, the challenge is ongoing. A combination of technical solutions, regulatory frameworks, and conscientious data management is required to continue making progress in this crucial area.

4.7 Future of Economics, Skills

The high probability of job automation, particularly in roles characterized by repetitive tasks and ordinary skills, is a significant concern for the future economy. While it is likely that new job opportunities emerge, it is anticipated that existing positions may be eliminated at a faster rate than new ones are created (West, 2018). This could potentially lead to a net decrease in total employment. *"There is a risk of unemployment and underemployment owing to the increased use of robotics and AI, leading potentially to inequality on a massive scale. The market on its own cannot resolve this knot of issues"* (West, 2018, p. 157).

Employment dynamics are expected to become more fluid, with individuals likely changing professions more frequently. *"In the Industry 4.0 context, employees need to have access to a new, improved, and unique lifelong education system so that they are properly educated for future jobs*" (Cramarenco et al., 2023, p. 736). Reskilling those whose jobs are automated, particularly those with low digital skills, and upskilling highly skilled workers, will be essential to mitigate misemployment and adapt to these changes (Cramarenco et al., 2023).

However, the rise in short-term jobs poses challenges to traditional methods of accruing social benefits, such as health and retirement, in many regions. This issue, combined with facing the loss of many positions manifest the urgency of political and social adjustments to redefine work and prevent social inequality. One proposal is to expand the definition of work to include socially beneficial activities like volunteering, parenting, and mentoring, compensating these tasks accordingly.

Even with inclusion of such approaches, the future of work remains uncertain. One widely discussed and more extensive approach is the introduction of an UBI, where all citizens receive a basic income sufficient for living, regardless of their employment status. While working would still provide additional financial means, an UBI could serve as a safety net in a highly automated economy.

In a future where a significant portion of work is automated and appropriate political and social adjustments are made, the role of jobs in people's lives might evolve. Society could return to a historical lifestyle where work is a part of one's identity but not the entirety of existence. This shift could allow more time for non-work activities, including art, culture, music, sports, and theater, fostering opportunities for personal enrichment. The economy might transition from a focus on consumption to one of creativity and self-expression, potentially leading to a more fulfilled and happier society (West, 2018).

4.8 Power Distribution and Monopoly Prevention

The landscape of AI innovation has recently seen a significant concentration of advancements within major tech companies. These firms, such as Microsoft and Google, have distinct advantages, particularly in terms of computational power and data access. This concentration of resources has made it challenging for academia and open-source projects to compete. Microsoft and Google are behind widely used models like ChatGPT and Bard. Without regulatory intervention, there is a risk that monopolies could form in a field as crucial and influential as AI, presenting significant risks (Bommasani et al., 2022).

To counteract this trend, it is suggested that governments step up and intervene to level existing advantages. Providing resources to academic institutions could enable them to compete more effectively in the field of AI development. *"To truly 'democratize' AI, a concerted effort by policymakers, academic institutions, and firm-level actors is needed to tackle the compute divide"* (Ahmed & Wahed, 2020, p. 39). This approach would not only foster innovation but also ensure a more diverse and competitive AI landscape (Ahmed & Wahed, 2020).

In addition to supporting academia, there is a growing call for stronger regulations targeting big tech companies, in general, as they become increasingly prominent in many aspects of daily life. These firms already inherit considerable power and regulating them could involve several measures.

One approach could be to enforce greater transparency regarding their data practices and usage. Another proposal is the regulation of mergers and acquisitions, as seen in the case of companies like OpenAI and Microsoft. Some advocates go further, proposing the breakup of these large companies to prevent them from becoming overly dominant and powerful (Richter et al., 2021; Wörsdörfer, 2022).

4.9 Reducing Disinformation

Reducing the disinformation potentially induced by generative AI is a high priority, especially considering the risks that scaled disinformation campaigns could pose to democracy. The primary focus here is on text-based generative models, as text is the most prevalent medium for information transmission. Secondary, other modalities, apart from deepfakes, are not as thoroughly explored in terms of their disinformation potential (Vaccari & Chadwick, 2020).

Developing more truthful and factually accurate AI systems is one approach to combating disinformation. Current research indicates the possibility of creating such systems. Curated datasets could potentially be an aid in developing "truthful" AI (Evans et al., 2021). Giving a LLM access to the web has been found to increase the factuality of its output, too (Nakano et al., 2022).

However, achieving a holistic state of truth-integrated AI technology in society faces significant barriers. Firstly, clear truthfulness standards need to be identified. Additionally, the creation of institutions to judge compliance with these standards both before and after deployment would be required. Biases in what is considered "true" or "false" are likely to persist, as well. *"Even a language model that produced no false claims could still be used to produce politically slanted or unfalsifiable statements"* (Goldstein et al., 2023, p. 44). Therefore, even with the development of optimal truthful AI, its impact may be limited to certain types of influence operations, though remaining a desirable goal. Still, the amount and quality of coordination required between AI developers as well as regulatory bodies would be immense (Goldstein et al., 2023).

Actions to maintain the distinguishability of synthetic content would also help to mitigate the risks of AI-based disinformation. Recognizing content as AI-generated, particularly in personalized and automated campaigns, could reduce its potential harm (Goldstein et al., 2023). However, since both malicious actors and regular entities use AI tools to create content, recognizability of synthetic content does not necessarily equate to immunity against disinforming synthetic content. But clarifying whether content, especially visual content like deepfakes, is artificial could help reduce misinformation (Vaccari & Chadwick, 2020).

Many doubt that there will be holistic mitigations capable of identifying if a message had an automated author, albeit there are various approaches research is conducted on. The retrospective detection of short textual AI content, commonly used in disinformation campaigns, might lack clear detectability due to limited complexity.

Tackling this problem at another level, such as limiting the scale at which these operations unfold, could be more effective. The primary advantage of these techniques is scalability due to automation. Disinformation campaigns require infrastructure, including inauthentic accounts and dissemination platforms. *"The best mitigation for automated content generation in disinformation thus is not to focus on the content itself, but on the infrastructure that distributes that content"* (Ben Buchanan et al., 2021, p. 45). In accordance with this revelation, regulatory efforts to reduce fake accounts could be intensified (Ben Buchanan et al., 2021).

Requiring proof of personhood, as proposed before, could substantially reduce the power of malicious actors. Furthermore, engaging institutions in media literacy campaigns can help mitigate risks. Raising awareness and providing knowledge about these topics can lead to more informed public discourse and incentivize further research. While this might reduce public trust, it is crucial for the public to be aware of the facts (Goldstein et al., 2023).

5 Future Outlook

The capability of generative AI to shape our future has become clear. While the possibilities are vast, a consensus among AI researchers is the inherent uncertainty surrounding AI's trajectory (Grace et al., 2024). Historical attempts at forecasting the future of any area, including the future of generative AI, caution against making overly precise predictions, favoring a more general outlook based on current trends and expert forecasts (Løhre & Teigen, 2017).

The pace of AI development, particularly in generative AI, shows no signs of decreasing. On the contrary, it is accelerating, with milestone achievements being reached sooner than previously anticipated and forecasts for prospective achievements becoming earlier (X. Tang et al., 2020). For instance, while 2022 predictions estimated a ten percent chance of AI outperforming humans in all tasks by 2037, in 2023, this forecast had shifted to as early as 2027. Furthermore, the likelihood of reaching this accomplishment by 2046 is now seen as fifty percent (Grace et al., 2024). This rapid advancement underscores the transformative potential of AI across society, yet it also highlights the ambivalent nature of such technologies.

Advanced AI systems inherit the potential for both extraordinary benefits and significant risks. Approximately forty-five percent of AI researchers acknowledge a non-negligible risk of catastrophic outcomes, as severe as human extinction. Despite these concerns, the majority remains rather optimistic, believing in the greater likelihood of AI inducing net positive rather than negative outcomes for humanity. This optimism, however, is accompanied by a uniform agreement that safety research has to prioritized more adequately to mitigate the considerable risks associated with AI's rapid development (Grace et al., 2024). The hurried competitive landscape among AI developers, focused to push the previous limits, may cause compromises regarding the thoroughness of safety evaluations (Scharre, 2021). Therefore, potentially less safe systems may be deployed, especially when the roles inside of the market are not yet clearly established.

There is an evident necessity for robust regulation, which is substantial in guiding the development of AI technologies in a manner that safeguards public interest and security (Luckett, 2023).

A topic beyond the scope of this thesis, yet of primary importance, is the concept of superhuman AI—an intelligence that surpasses human capabilities in all respects. The emergence of such AI could fundamentally alter society, potentially inducing unprecedented advancements in science and technology. Whilst philosophers, scientists and authors have been imagining the consequences of a second highly intelligent entity for a long time, the exact nature of these changes remains beyond our current capacity to predict, as human cognitive limits would be exceeded (Bostrom, 2017; Diederich, 2021; S. Russell, 2017).

The collective efforts of researchers, policymakers, and society at large will be essential in aligning AI development towards a future that maximizes benefits while mitigating risks. The future of AI, as with any revolutionary technology, will be shaped not only by its creators but from various actors across the whole bandwidth of society.

6 Conclusion

The exploration of generative AI has illuminated its capacity to transform various facets of society. This potential includes both positive advancements and negative implications. The examination has underscored the necessity for a more proactive engagement with the challenges presented by generative AI, as passive or inadequate responses could induce additional harm. Current mitigation strategies, while numerous, often lack comprehensiveness or depend on voluntary participation, highlighting the need for structured regulatory frameworks. Regulatory bodies may be required to establish and periodically update dedicated guidelines and rules than previously introduced to ensure the responsible development and implementation of these technologies.

Furthermore, the importance of collaborative efforts among diverse stakeholders in navigating the ethical and societal implications of generative AI is highlighted. As no technology is flawless, collective efforts for safety and ethical evaluations are required, rather than leaving these considerations to individual entities. Such collaboration is essential, given the high stakes associated with the deployment of advanced AI systems.

In conclusion, for achieving an effective and responsible integration of generative AI in various operations, a wide array of limitations and considerations must be overcome. But when executed successfully, and an environment of oversight, continuous dialogue, and collaborative problem-solving is promoted, society can steer the development of generative AI towards outcomes that are beneficial and aligned with societal values.

7 Limitations

This thesis encounters several limitations that necessitate acknowledgment.

Firstly, the extensive range of topics covered means that in-depth analysis for certain areas is constrained, and in some cases, topics may not be addressed at all. As a result, some subjects are only briefly examined, providing an overview rather than a detailed exploration. This breadth of coverage inevitably leads to a loss of detail in some discussions.

Another limitation is the rapidly evolving nature of the field of generative AI. With ongoing research and continuous advancements, some of the content presented may quickly become outdated or may not fully capture the most recent developments.

Because the focus is on the negative implications of generative AI and the strategies for mitigating these risks, the significant positive impacts and benefits of these technologies may be underrepresented. This focus could skew the overall perspective presented in the thesis.

Moreover, due to intersections of areas and imbalance in current research states, not every examined aspect of ethical and societal implications has its corresponding regulatory framework or control mechanism addressed. This discrepancy may leave some areas less thoroughly explored than others.

The thesis does not always make a clear distinction between AI in general and generative AI, as these areas are often closely intertwined. Some issues discussed, such as the lack of transparency, are not exclusive to generative AI. This lack of clear separation could lead to confusion about the applicability of certain implications and strategies.

Additionally, while the thesis title suggests an ethical evaluation, the focus is more on tangible aspects of generative AI, with ethical considerations being secondary rather than being the primary focus of analysis.

Acknowledging these limitations is important for a balanced understanding of the thesis's scope and the areas where further research is needed.

7 Bibliography

Abid, A., Farooqi, M., & Zou, J. (2021). *Persistent Anti-Muslim Bias in Large Language Models* (arXiv:2101.05783). arXiv. https://doi.org/10.48550/arXiv.2101.05783

Adobe. (2023). *Adobe Firefly—Free Generative AI for creatives*. Adobe Firefly - Free Generative AI for Creatives. https://www.adobe.com/products/firefly.html

Ahmed, N., & Wahed, M. (2020). *The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research* (arXiv:2010.15581). arXiv. https://doi.org/10.48550/arXiv.2010.15581

Ahmed, N., Wahed, M., & Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, *379*(6635), 884–886. https://doi.org/10.1126/science.ade2420

Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: The risk of rationalization. *Proceedings of the 36th International Conference on Machine Learning*, 161–170. https://proceedings.mlr.press/v97/aivodji19a.html

Arik, S., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural Voice Cloning with a Few Samples. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper_files/paper/2018/hash/4559912e7a94a9c32b09d894f2bc3c82-Abstract.html

Ben Buchanan, Andrew Lohn, Micah Musser, & Katerina Sedova. (2021). *Truth, Lies, and Automation: How Language Models Could Change Disinformation* (p. 53). Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/truth-lies-and-automation/

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bergman, C. (2022, December). The AI content flood. *Nieman Lab*. https://www.niemanlab.org/2022/12/the-ai-content-flood/

Berry, D. M. (2023). The Limits of Computation: Joseph Weizenbaum and the ELIZA Chatbot. *Weizenbaum Journal of the Digital Society*, *3*(3), Article 3. https://doi.org/10.34669/WI.WJDS/3.3.2

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504. https://doi.org/10.1145/3593013.3594095

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* (arXiv:1607.06520). arXiv. https://doi.org/10.48550/arXiv.1607.06520

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. http://arxiv.org/abs/2108.07258

Bostrom, N. (2017). *Superintelligence: Paths, dangers, strategies* (Reprinted with corrections 2017). Oxford University Press.

Boxin Wang, Bo Li, & Zinan Lin. (2023, October 16). DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *Microsoft Research*. https://www.microsoft.com/en-us/research/blog/decodingtrust-a-comprehensive-assessment-of-trustworthiness-in-gpt-models/

Brittain, B. (2023, September 6). US Copyright Office denies protection for another AI-created image. *Reuters*. https://www.reuters.com/legal/litigation/us-copyright-office-denies-protection-another-ai-created-image-2023-09-06/

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712

Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., Nogues, I., Li, X., Guo, W., Wang, Y., Fang, W., Qiu, M., Hou, Y., Kovarnik, T., Vocka, M., … Lu, J. (2023). Large-scale pancreatic cancer detection via non-contrast CT and deep learning. *Nature Medicine*, 1–11. https://doi.org/10.1038/s41591-023-02640-w

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT* (arXiv:2303.04226). arXiv. https://doi.org/10.48550/arXiv.2303.04226

Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. (2021). Deepfake: An Overview. In P. K. Singh, S. T. Wierzchoń, S. Tanwar, M. Ganzha, & J. J. P. C. Rodrigues (Eds.), *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security* (pp. 557–566). Springer. https://doi.org/10.1007/978-981-16-0733-2_39

Chang, K. K., Cramer, M., Soni, S., & Bamman, D. (2023). *Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4* (arXiv:2305.00118). arXiv. https://doi.org/10.48550/arXiv.2305.00118

Chen, C., Li, Y., Wu, Z., Xu, M., Wang, R., & Zheng, Z. (2023). Towards Reliable Utilization of AIGC: Blockchain-Empowered Ownership Verification Mechanism. *IEEE Open Journal of the Computer Society*, *4*, 326–337. https://doi.org/10.1109/OJCS.2023.3315835

Copyright, Designs and Patents Act 1988, Pub. L. No. c. 48, Part I, Chapter I (1988). https://www.legislation.gov.uk/ukpga/1988/48/part/I/chapter/I/crossheading/authorship-and-ownership-of-copyright/2013-11-01

Cramarenco, R. E., Burcă-Voicu, M. I., & Dabija, D. C. (2023). The impact of artificial intelligence (AI) on employees' skills and well-being in global labor markets: A systematic review. *Oeconomia Copernicana*, *14*(3), 731–767. https://doi.org/10.24136/oc.2023.022

Cuccovillo, L., Papastergiopoulos, C., Vafeiadis, A., Yaroshchuk, A., Aichroth, P., Votis, K., & Tzovaras, D. (2022). Open Challenges in Synthetic Speech Detection. *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6. https://doi.org/10.1109/WIFS55849.2022.9975433

Dahmani, L., & Bohbot, V. D. (2020). Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-62877-0

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep37825

Diederich, J. (2021). *The Psychology of Artificial Superintelligence* (Vol. 42). Springer International Publishing. https://doi.org/10.1007/978-3-030-71842-8

Dodig-Crnkovic, G. (2023). How GPT Realizes Leibniz's Dream and Passes the Turing Test without Being Conscious. *Computer Sciences & Mathematics Forum*, *8*(1), Article 1. https://doi.org/10.3390/cmsf2023008066

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology*, *13*, 100060. https://doi.org/10.1016/j.jrt.2023.100060

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*(1), Article 1. https://doi.org/10.1007/s40979-023-00140-5

Emmert-Streib, F., Yli-Harja, O., & Dehmer, M. (2020). Artificial Intelligence: A Clarification of Misconceptions, Myths and Desired Status. *Frontiers in Artificial Intelligence*, *3*. https://www.frontiersin.org/articles/10.3389/frai.2020.524339

European Comission. (2018, December 18). *A definition of Artificial Intelligence: Main capabilities and scientific disciplines | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

European Commission. (2023, December 9). *Commission welcomes political agreement on AI Act* [Press Release]. Commission Welcomes Political Agreement on Artificial Intelligence Act. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473

European Union Agency for Law Enforcement Cooperation. (2022). *Facing reality?: Law enforcement and the challenge of deepfakes: An observatory report from the Europol innovation lab.* Publications Office. https://data.europa.eu/doi/10.2813/08370

Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., Righetti, L., & Saunders, W. (2021). *Truthful AI: Developing and governing AI that does not lie* (arXiv:2110.06674). arXiv. http://arxiv.org/abs/2110.06674

Fan, Y., Xie, M., Wu, P., & Yang, G. (2022). Real-Time Deepfake System for Live Streaming. *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 202–205. https://doi.org/10.1145/3512527.3531350

Gerlich, M. (2023). The Power of Virtual Influencers: Impact on Consumer Behaviour and Attitudes in the Age of AI. *Administrative Sciences*, *13*(8), Article 8. https://doi.org/10.3390/admsci13080178

Gillotte, J. (2020). *Copyright Infringement in AI-Generated Artworks* (SSRN Scholarly Paper 3657423). https://papers.ssrn.com/abstract=3657423

Goldman Sachs Research. (2023, April 5). *Generative AI Could Raise Global GDP by 7%*. Goldman Sachs. https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations* (arXiv:2301.04246). arXiv. http://arxiv.org/abs/2301.04246

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a 'right to explanation'. *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Göring, S., Ramachandra Rao, R. R., Merten, R., & Raake, A. (2023). Analysis of Appeal for Realistic AI-Generated Photos. *IEEE Access*, *11*, 38999–39012. https://doi.org/10.1109/ACCESS.2023.3267968

Gowal, S., & Kohli, P. (2023, August 29). Identifying AI-generated images with SynthID. *Google DeepMind*. https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/

Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). *A survey of Generative AI Applications* (arXiv:2306.02781). arXiv. https://doi.org/10.48550/arXiv.2306.02781

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). *Thousands of AI Authors on the Future of AI* (arXiv:2401.02843). arXiv. http://arxiv.org/abs/2401.02843

Grassini, S. (2023). Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Education Sciences*, *13*(7), Article 7. https://doi.org/10.3390/educsci13070692

Gujar, S., Shah, T., Honawale, D., Bhosale, V., Khan, F., Verma, D., & Ranjan, R. (2022). GenEthos: A Synthetic Data Generation System With Bias Detection And Mitigation. *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, 1–6. https://doi.org/10.1109/IC3SIS54991.2022.9885653

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). *Textbooks Are All You Need* (arXiv:2306.11644). arXiv. http://arxiv.org/abs/2306.11644

Gupta, M., Malik, T., & Sinha, C. (2022). Delivery of a Mental Health Intervention for Chronic Pain Through an Artificial Intelligence–Enabled App (Wysa): Protocol for a Prospective Pilot Study. *JMIR Research Protocols*, *11*(3), e36910. https://doi.org/10.2196/36910

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123. https://doi.org/10.1145/3593013.3594067

Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*, 000812561986492. https://doi.org/10.1177/0008125619864925

Haresamudram, K., Larsson, S., & Heintz, F. (2023). Three Levels of AI Transparency. *Computer*, *56*(2), 93–100. https://doi.org/10.1109/MC.2022.3213181

Hinds, J., Williams, E. J., & Joinson, A. N. (2020). “It wouldn’t happen to me”: Privacy concerns and perspectives following the Cambridge Analytica scandal. *International Journal of Human-Computer Studies*, *143*, 102498. https://doi.org/10.1016/j.ijhcs.2020.102498

Hosseini, D. (2023, August 8). *Generative AI: a problematic illustration of the intersections of racialized gender, race, ethnicity*. Dustin Hosseini. https://www.dustinhosseini.com/blog/2023/08/08/generative-ai-a-problematic-illustration-of-the-intersections-of-racialized-gender-race-ethnicity

Huang, H., Sun, N., & Lin, X. (2022). Blockwise Spectral Analysis for Deepfake Detection in High-fidelity Videos. *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–9. https://doi.org/10.1109/DSAA54385.2022.10032370

Huang, S., & Siddarth, D. (2023). *Generative AI and the Digital Commons* (arXiv:2303.11074). arXiv. http://arxiv.org/abs/2303.11074

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). *Social Biases in NLP Models as Barriers for Persons with Disabilities* (arXiv:2005.00813). arXiv. https://doi.org/10.48550/arXiv.2005.00813

IBM Data and AI Team. (2023, July 6). AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What’s the difference? *IBM Blog*. https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks

Jacob Ridley. (2023, September 21). OpenAI's new DALL-E 3 AI image generator isn't allowed to copy a living artist's style by name. *PC Gamer*. https://www.pcgamer.com/openais-new-dall-e-3-ai-image-generator-isnt-allowed-to-copy-a-living-artists-style-by-name/

Jan Hatzius, Joseph Briggs, Devesh Kodnani, & Giovanni Pierdomenico. (2023). *The Potentially Large Effects of Artificial Intelligence on Economic Growth*.

Kahn, C. E. (2022). Hitting the Mark: Reducing Bias in AI Systems. *Radiology: Artificial Intelligence*, *4*(5), e220171. https://doi.org/10.1148/ryai.220171

Kapur, S. (2021). Reducing racial bias in AI models for clinical use requires a top-down intervention. *Nature Machine Intelligence*, *3*(6), Article 6. https://doi.org/10.1038/s42256-021-00362-7

Kerbl, B., Kopanas, G., Leimkuehler, T., & Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, *42*(4), 139:1-139:14. https://doi.org/10.1145/3592433

LaGrandeur, K. (2023). The consequences of AI hype. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00352-y

Lee, P. Y. K., Ma, N. F., Kim, I.-J., & Yoon, D. (2023). Speculating on Risks of AI Clones to Selfhood and Relationships: Doppelganger-phobia, Identity Fragmentation, and Living Memories. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 91:1-91:28. https://doi.org/10.1145/3579524

Li, H. Y., An, J. T., & Zhang, Y. (2021). Ethical Problems and Countermeasures of Artificial Intelligence Technology. *E3S Web of Conferences*, *251*, 01063. https://doi.org/10.1051/e3sconf/202125101063

Library of Congress, Copyright Office. (2023). *Artificial Intelligence and Copyright* (Docket No. 2023-6; pp. 59942–59949). https://www.federalregister.gov/documents/2023/08/30/2023-18624/artificial-intelligence-and-copyright

Liu, J. (2022). Deep learning algorithms driven by artificial intelligence technologies may cause the disruption of information interaction mechanisms and the creation of technological monopolies. *International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2022)*, *12256*, 413–419. https://doi.org/10.1117/12.2635677

Løhre, E., & Teigen, K. H. (2017). Probabilities associated with precise and vague forecasts. *Journal of Behavioral Decision Making*, *30*(5), 1014–1026. https://doi.org/10.1002/bdm.2021

Lola Mayor. (2023, July 20). AI: Digital artist's work copied more times than Picasso. *BBC News*. https://www.bbc.com/news/uk-wales-66099850

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, *7*(1), Article 1. https://doi.org/10.1038/s41562-022-01460-1

Luckett, J. (2023). Regulating Generative AI: A Pathway to Ethical and Responsible Implementation. *International Journal on Cybernetics & Informatics*, *12*(5), 79–92. https://doi.org/10.5121/ijci.2023.120508

Lucy, L., & Bamman, D. (2021). Gender and Representation Bias in GPT-3 Generated Stories. In N. Akoury, F. Brahman, S. Chaturvedi, E. Clark, M. Iyyer, & L. J. Martin (Eds.), *Proceedings of the Third Workshop on Narrative Understanding* (pp. 48–55). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nuse-1.5

Mahajan, D., Gapat, A., Moharkar, L., Sawant, P., & Dongardive, K. (2021). Artificial Generation of Realistic Voices. *International Journal of Applied Sciences and Smart Technologies*, *3*(1), Article 1. https://doi.org/10.24071/ijasst.v3i1.2744

Mezrich, J. L. (2022). Is Artificial Intelligence (AI) a Pipe Dream? Why Legal Issues Present Significant Hurdles to AI Autonomy. *American Journal of Roentgenology*, *219*(1), 152–156. https://doi.org/10.2214/AJR.21.27224

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman, J. (2022). *WebGPT: Browser-assisted question-answering with human feedback* (arXiv:2112.09332). arXiv. https://doi.org/10.48550/arXiv.2112.09332

Nalini, D. C., Kumar, D. R. A., & Professor, A. (2023). *Generative AI: A Comprehensive Study of Advancements and Application*. https://www.semanticscholar.org/paper/Generative-AI%3A-A-Comprehensive-Study-of-and-Nalini-Kumar/70cb05a3179bb1c8b65891dd8b7a4add1fe0eb2b

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., & Lee, K. (2023). *Scalable Extraction of Training Data from (Production) Language Models* (arXiv:2311.17035). arXiv. https://doi.org/10.48550/arXiv.2311.17035

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence - CHI '94*, 72–78. https://doi.org/10.1145/191666.191703

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 1–6. https://doi.org/10.1109/ICTKE.2017.8259629

OpenAI. (2022, November 30). Introducing ChatGPT. *Introducing ChatGPT*. https://openai.com/blog/chatgpt

OpenAI. (2024, January 10). Introducing the GPT Store. *OpenAI Blog*. https://openai.com/blog/introducing-the-gpt-store

OpenAI Reseach. (2021, January 5). *DALL·E: Creating images from text*. https://openai.com/research/dall-e

OpenAI Reseach. (2024). *Video generation models as world simulators*. https://openai.com/research/video-generation-models-as-world-simulators

Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*, *23*(5), 719–735. https://doi.org/10.1080/1369118X.2020.1713842

Pernias, P., Rampas, D., Richter, M. L., Pal, C. J., & Aubreville, M. (2023). *Wuerstchen: An Efficient Architecture for Large-Scale Text-to-Image Diffusion Models* (arXiv:2306.00637). arXiv. http://arxiv.org/abs/2306.00637

Pischedda, D., Kaufmann, V., Wudarczyk, O. A., Rahman, R. A., Hafner, V. V., Kuhlen, A. K., & Haynes, J.-D. (2023). Human or AI? The brain knows it! A brain-based Turing Test to discriminate between human and artificial agents. *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 951–958. https://doi.org/10.1109/RO-MAN57019.2023.10309541

Poba-Nzaou, P., Galani, M., Uwizeyemungu, S., & Ceric, A. (2021). The impacts of artificial intelligence (AI) on jobs: An industry perspective. *Strategic HR Review*, *20*(2), 60–65. https://doi.org/10.1108/SHR-01-2021-0003

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rampas, D., Pernias, P., & Aubreville, M. (2023). *A Novel Sampling Scheme for Text- and Image-Conditional Image Synthesis in Quantized Latent Spaces* (arXiv:2211.07292). arXiv. http://arxiv.org/abs/2211.07292

Richter, H., Straub, M., Tuchtfeld, E., Buri, I., van Hoboken, J., De Gregorio, G., Pollicino, O., Peukert, A., Appelman, N., Quintais, J. P., Fahy, R., Zech, H., Goanta, C., Ruschemeier, H., Leerssen, P., Janal, R., Rodríguez de las Heras Ballell, T., Graef, I., Franck, J.-U., … Vergnolle, D. S. (2021). *To Break Up or Regulate Big Tech? Avenues to Constrain Private Power in the DSA/DMA Package* (SSRN Scholarly Paper 3932809). https://doi.org/10.2139/ssrn.3932809

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1), Article 1. https://doi.org/10.37074/jalt.2023.6.1.9

Russell, S. (2017). Artificial intelligence: The future is superintelligent. *Nature*, *548*(7669), Article 7669. https://doi.org/10.1038/548520a

Russell, S. J., Norvig, P., Chang, M., Devlin, J., Dragan, A., Forsyth, D., Goodfellow, I., Malik, J., Mansinghka, V., Pearl, J., & Wooldridge, M. J. (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.

Sablayrolles, A., Douze, M., Schmid, C., & Jegou, H. (2020). Radioactive data: Tracing through training. *Proceedings of the 37th International Conference on Machine Learning*, 8326–8335. https://proceedings.mlr.press/v119/sablayrolles20a.html

Scharre, P. (2021). *Debunking the AI Arms Race Theory (Summer 2021)*. https://hdl.handle.net/2152/87035

Schick, N. (2020). *Deepfakes: The coming infocalypse* (First U.S. edition). Twelve.

Selenko, E., Bankins, S., Shoss, M., Warburton, J., & Restubog, S. L. D. (2022). Artificial Intelligence and the Future of Work: A Functional-Identity Perspective. *Current Directions in Psychological Science*, *31*(3), 272–279. https://doi.org/10.1177/09637214221091823

Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S. Y., Fang, P.-H., & Ming, W.-K. (2019). Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. *JMIR Medical Informatics*, *7*(3), e10010. https://doi.org/10.2196/10010

Sun, J., Li, M., Chen, Z., Zhang, Y., Wang, S., & Moens, M.-F. (2023). *Contrast, Attend and Diffuse to Decode High-Resolution Images from Brain Activities* (arXiv:2305.17214). arXiv. https://doi.org/10.48550/arXiv.2305.17214

Sundar Pichai. (2023, April 17). *Google CEO: AI impact to be more profound than discovery of fire, electricity | 60 Minutes* (Scott Pelley, Interviewer) [Interview]. https://www.youtube.com/watch?v=W6HpE1rhs7w

Tang, L., Ruiz, N., Chu, Q., Li, Y., Holynski, A., Jacobs, D. E., Hariharan, B., Pritch, Y., Wadhwa, N., Aberman, K., & Rubinstein, M. (2023). *RealFill: Reference-Driven Generation for Authentic Image Completion* (arXiv:2309.16668). arXiv. https://doi.org/10.48550/arXiv.2309.16668

Tang, X., Li, X., Ding, Y., Song, M., & Bu, Y. (2020). The pace of artificial intelligence innovations: Speed, talent, and trial-and-error. *Journal of Informetrics*, *14*(4), 101094. https://doi.org/10.1016/j.joi.2020.101094

Tate, R.-M. (2023, May 15). *Catching bad content in the age of AI* [Article]. MIT Technology Review. https://www.technologyreview.com/2023/05/15/1073019/catching-bad-content-in-the-age-of-ai/

Taulli, T. (2023). *Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business*. Apress. https://doi.org/10.1007/978-1-4842-9367-6

The Act | The Artificial Intelligence Act (2021). https://artificialintelligenceact.eu/the-act/

Trevithick, A., Chan, M., Stengel, M., Chan, E. R., Liu, C., Yu, Z., Khamis, S., Chandraker, M., Ramamoorthi, R., & Nagano, K. (2023). *Real-Time Radiance Fields for Single-Image Portrait View Synthesis* (arXiv:2305.02310). arXiv. http://arxiv.org/abs/2305.02310

Turner, J. (2019). *Robot Rules: Regulating Artificial Intelligence*. Springer International Publishing. https://doi.org/10.1007/978-3-319-96235-1

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

Wang, X., Thakker, M., Chen, Z., Kanda, N., Eskimez, S. E., Chen, S., Tang, M., Liu, S., Li, J., & Yoshioka, T. (2023). *SpeechX: Neural Codec Language Model as a Versatile Speech Transformer* (arXiv:2308.06873). arXiv. https://doi.org/10.48550/arXiv.2308.06873

Watiktinnakorn, C., Seesai, J., & Kerdvibulvech, C. (2023). Blurring the lines: How AI is redefining artistic ownership and copyright. *Discover Artificial Intelligence*, *3*(1), 37. https://doi.org/10.1007/s44163-023-00088-y

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). *Testing of Detection Tools for AI-Generated Text* (arXiv:2306.15666). arXiv. https://doi.org/10.48550/arXiv.2306.15666

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). *Sociotechnical Safety Evaluation of Generative AI Systems* (arXiv:2310.11986). arXiv. http://arxiv.org/abs/2310.11986

West, D. M. (2018). *The future of work: Robots, AI, and automation*. Brookings Institution Press.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11). https://timreview.ca/sites/default/files/article_PDF/TIMReview_November2019%20-%20D%20-%20Final.pdf

World Bank. (2023). *Identification for Development (ID4D) and Digitalizing G2P Payments (G2Px) 2022 Annual Report* (Annual Report 179817; ID4D/G2Px). World Bank Group. https://documents1.worldbank.org/curated/en/099437402012317995/pdf/IDU00fd54093061a70475b0a3b50dd7e6cdfe147.pdf

Wörsdörfer, M. (2022). What Happened to 'Big Tech' and Antitrust? And How to Fix Them! *Philosophy of Management*, *21*(3), 345–369. https://doi.org/10.1007/s40926-022-00193-5

Xie, T., Liao, L., Bi, C., Tang, B., Yin, X., Yang, J., Wang, M., Yao, J., Zhang, Y., & Ma, Z. (2021). Towards Realistic Visual Dubbing with Heterogeneous Sources. *Proceedings of the 29th ACM International Conference on Multimedia*, 1739–1747. https://doi.org/10.1145/3474085.3475318

Yildirim, A. B., Pehlivan, H., Bilecen, B. B., & Dundar, A. (2023). *Diverse Inpainting and Editing with GAN Inversion* (arXiv:2307.15033). arXiv. https://doi.org/10.48550/arXiv.2307.15033

Zdravkova, K., Krasniqi, V., Dalipi, F., & Ferati, M. (2022). Cutting-edge communication and learning assistive technologies for disabled children: An artificial intelligence perspective. *Frontiers in Artificial Intelligence*, *5*. https://www.frontiersin.org/articles/10.3389/frai.2022.970430

Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R. J., Jia, Y., Rosenberg, A., & Ramabhadran, B. (2019). *Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning* (arXiv:1907.04448). arXiv. http://arxiv.org/abs/1907.04448

Zhong, H., Chang, J., Yang, Z., Wu, T., Mahawaga Arachchige, P. C., Pathmabandu, C., & Xue, M. (2023). Copyright Protection and Accountability of Generative AI: Attack, Watermarking and Attribution. *Companion Proceedings of the ACM Web Conference 2023*, 94–98. https://doi.org/10.1145/3543873.3587321

Zhu, Y., Baca, J., Rekabdar, B., & Rawassizadeh, R. (2023). *A Survey of AI Music Generation Tools and Models* (arXiv:2308.12982). arXiv. http://arxiv.org/abs/2308.12982

9 List of Figures