In The Age of AI: Ethical, legal and social challenges

Master Thesis

For attainment of the academic degree of
Master of Science in Engineering (MSc)

Master Programme Cyber Security and Resilience
at **St. Pölten University of Applied Sciences**

by
Miss Maryam Al-Durra
Cr212501

First advisor: Simon Tjoa

Sankt Pölten 19th September 2023

Declaration

I hereby affirm that

- I have written this thesis independently, that I have not used any sources or aids other than those indicated, and that I have not made use of any unauthorised assistance.
- I have not previously submitted this thesis topic to an assessor, either in Austria or abroad, for evaluation or as an examination paper in any form.
- This thesis corresponds to the thesis assessed by the assessor.

.... Sankt Pölten 19/09/2023.........   ...Maryam Al-Durra.................

Place, Date   Signature

Preface

The rapid pace of AI implementation has brought about significant transformations in society. The primary objective of artificial intelligence (AI) has been to bring about a transformative impact on various industries by developing intelligent machines capable of executing tasks that conventionally necessitate human intelligence. Artificial intelligence (AI) aims to replicate human cognitive functions, including perception, analysis, problem-solving, and learning, with the overarching objective of achieving autonomous operation, adaptability to novel circumstances, and ongoing performance enhancement. In addition to the aforementioned advantages, artificial intelligence (AI) also presents a variety of ethical, legal, and social drawbacks that necessitate careful consideration of responsible governance.

The objective of this thesis is to analyze the various challenges presented by artificial intelligence (AI) through a multidisciplinary lens, with a specific emphasis on ethical, legal, and social dimensions. The present thesis investigates the various methodologies through which artificial intelligence (AI) can perpetuate biases, perpetrate injustices, and other ethical, legal and social challenges associated to AI. The objective of this study is to analyze the current legal framework governing artificial intelligence (AI) and propose suggestions for enhancing it in order to ensure responsible and transparent development and deployment of AI.

By conducting a comprehensive analysis of relevant scholarly works, engaging in interviews, and examining case studies involving experts in the field, this thesis offers a meticulous examination of the complexity introduced by artificial intelligence (AI). Furthermore, it presents valuable perspectives on the potential strategies that can be pursued to safeguard societal well-being. The discoveries and recommendations presented in this thesis have the potential to have substantial implications for policymakers, civil society organizations, leaders, and other entities involved in promoting responsible development and implementation of artificial intelligence.

Abstract

The expeditious progress of artificial intelligence (AI) has yielded notable advancements across diverse industries; however, it also presents a multitude of ethical, legal, and social dilemmas. In order to ensure the alignment of AI development with ethical principles, legal regulations, and social values, it is imperative to address these challenges with utmost attention.

This study undertakes a thorough examination of the complex obstacles posed by artificial intelligence (AI) by conducting a comprehensive analysis of ethical, legal, and social concerns. The initial focus of this discussion centres on the ethical dilemmas associated with artificial intelligence (AI), encompassing issues such as bias and transparency, privacy and security, as well as accountability and responsibility. This thesis assesses the efficacy of existing ethical frameworks and presents a more inclusive methodology to guarantee the development and utilization of AI in accordance with ethical principles.

Furthermore, this study examines the legal implications associated with artificial intelligence (AI), encompassing issues such as intellectual property rights, liability and responsibility, and regulatory frameworks. The assessment focuses on the sufficiency of current laws and regulations in effectively addressing the challenges. Additionally, it aims to pinpoint specific areas where legal reform is imperative to guarantee the development and utilization of AI aligns with established legal principles.

Finally, this analysis delves into the social implications of artificial intelligence (AI), specifically focusing on its effects on employment, inequality, and democracy. The assessment of potential social implications of AI and the formulation of strategies to address these challenges employ a multidisciplinary approach that integrates perspectives from economics, sociology, and political science. Moreover, given the paramount significance of cyber security in safeguarding the ethical implementation of artificial intelligence (AI), this thesis sheds light on potential attacks that pose a threat to the integrity of AI algorithms. Additionally, the study explores recommendations proposed in existing literature to mitigate these vulnerabilities and fortify the resilience of AI systems.

In its entirety, this thesis offers a thorough examination of the ethical, legal, and social complexities associated with artificial intelligence (AI). Additionally, it puts forth a set of approaches aimed at effectively addressing these challenges, with the ultimate goal of fostering the advancement of AI that is advantageous to the broader society.

Kurzfassung

Der rasante Fortschritt der künstlichen Intelligenz (KI) hat in verschiedenen Branchen zu bemerkenswerten Fortschritten geführt; Es bringt jedoch auch eine Vielzahl ethischer, rechtlicher und sozialer Dilemmata mit sich. Um die Ausrichtung der KI-Entwicklung an ethischen Grundsätzen, gesetzlichen Vorschriften und gesellschaftlichen Werten sicherzustellen, ist es zwingend erforderlich, diesen Herausforderungen mit größter Aufmerksamkeit zu begegnen.

In dieser Studie werden die komplexen Hindernisse, die künstliche Intelligenz (KI) mit sich bringt, gründlich untersucht, indem eine umfassende Analyse ethischer, rechtlicher und sozialer Bedenken durchgeführt wird. Der Schwerpunkt dieser Diskussion liegt zunächst auf den ethischen Dilemmata im Zusammenhang mit künstlicher Intelligenz (KI) und umfasst Themen wie Voreingenommenheit und Transparenz, Privatsphäre und Sicherheit sowie Rechenschaftspflicht und Verantwortung. Diese Arbeit bewertet die Wirksamkeit bestehender ethischer Rahmenwerke und stellt eine umfassendere Methodik vor, um die Entwicklung und Nutzung von KI im Einklang mit ethischen Grundsätzen zu gewährleisten.

Darüber hinaus untersucht diese Studie die rechtlichen Auswirkungen, die mit künstlicher Intelligenz (KI) verbunden sind, und umfasst Themen wie geistige Eigentumsrechte, Haftung und Verantwortung sowie regulatorische Rahmenbedingungen. Die Bewertung konzentriert sich darauf, ob die aktuellen Gesetze und Vorschriften ausreichen, um die oben genannten Herausforderungen wirksam zu bewältigen. Darüber hinaus zielt es darauf ab, bestimmte Bereiche zu ermitteln, in denen eine Rechtsreform unerlässlich ist, um sicherzustellen, dass die Entwicklung und Nutzung von KI im Einklang mit etablierten Rechtsgrundsätzen steht.

Abschließend befasst sich diese Analyse mit den sozialen Auswirkungen künstlicher Intelligenz (KI) und konzentriert sich dabei insbesondere auf ihre Auswirkungen auf Beschäftigung, Ungleichheit und Demokratie. Die Bewertung potenzieller sozialer Auswirkungen von KI und die Formulierung von Strategien zur Bewältigung dieser Herausforderungen basieren auf einem multidisziplinären Ansatz, der Perspektiven aus den Bereichen Wirtschaft, Soziologie und Politikwissenschaft integriert.

Angesichts der überragenden Bedeutung der Cybersicherheit für die ethische Umsetzung künstlicher Intelligenz (KI) beleuchtet diese Arbeit außerdem potenzielle Angriffe, die eine Bedrohung für die Integrität von KI-Algorithmen darstellen. Darüber hinaus untersucht die Studie in der vorhandenen Literatur vorgeschlagene Empfehlungen zur Minderung dieser Schwachstellen und zur Stärkung der Widerstandsfähigkeit von KI-Systemen.

Insgesamt bietet diese Arbeit eine gründliche Untersuchung der ethischen, rechtlichen und sozialen Komplexität, die mit künstlicher Intelligenz (KI) verbunden ist. Darüber hinaus werden eine Reihe von Ansätzen vorgestellt, die darauf abzielen, diese Herausforderungen wirksam anzugehen, mit dem ultimativen Ziel, die Weiterentwicklung der KI zu fördern, die für die Gesellschaft im Allgemeinen von Vorteil ist.

Table of Content

1 Introduction

Artificial Intelligence (AI) has emerged as a significant catalyst for transformative change, revolutionizing various industries and profoundly reshaping individual’s daily lives. This impact spans critical domains such as healthcare and humanitarian aid, as well as more commonplace areas like entertainment. The continuous development and advancement of artificial intelligence (AI) has resulted in its increasing integration into various domains within society, leading to the emergence of unique opportunities and enhanced efficiencies. Artificial intelligence (AI) possesses significant potential in fostering economic expansion, advancing social progress, and enhancing human welfare and safety. Nevertheless, it is important to acknowledge that this exponential expansion gives rise to numerous ethical, legal, and social dilemmas that necessitate thorough and ongoing scrutiny and contemplation.

The limited comprehensibility of AI-based technology, along with concerns regarding the security and privacy of data, as well as ethical considerations, present notable hazards to users, developers, and governmental entities. A crucial concern that arises with the advancement of artificial intelligence (AI) pertains to the ethical and moral dilemmas that accompany its development. The focus of this study will be on the ethical, social and moral implications that may be attributed to artificial intelligence (AI) or emerge as a consequence of its implementation.

This paper is aiming to answer the following questions :

- To what degree do current legal frameworks and scholarly investigations sufficiently encompass the ethical considerations that arise from the implementation of artificial intelligence (AI) technologies?
- What legislative measures or policy enhancements are required to ensure a more equitable and responsible integration of artificial intelligence (AI) into society?

- What are The impact of AI-deployed technologies on individual‘s daily lives in social, ethical, and legal dimensions.

The swift progress of Artificial Intelligence (AI) technologies in contemporary times has introduced a novel framework of ingenuity and metamorphosis in various industries, encompassing healthcare, finance, education, and entertainment. The advent of deep learning, natural language processing, and computer vision has significantly transformed our perception and interaction with machines, exemplifying a range of artificial intelligence capabilities. Nevertheless, the transformative potential of artificial intelligence (AI) is accompanied by a myriad of intricate ethical, social, and legal concerns that necessitate meticulous deliberation and thorough examination. Despite these concerns, Gary Marcus, a professor of cognitive science, summarizes ten limitations of deep learning namely*, “ it is data-hungry, it has limited capacity for transfer, it has no natural way to deal with hierarchical structure, it struggles with open-ended inference, it is not sufficiently transparent, it has not been well integrated with prior knowledge, it cannot inherently distinguish causation from correlation, it presumes a largely stable world, in ways that may be problematic,it works well as an approximation, but its answers often cannot be fully trusted, it is difficult to engineer with”* [60].

The integration of artificial intelligence (AI) into our everyday activities has given rise to significant ethical considerations pertaining to autonomy, accountability, and bias. The increasing autonomy and decision-making capabilities of AI systems necessitate a heightened consideration of the ethical implications associated with their actions. The complexity of accountability is heightened when AI systems rely on intricate algorithms that are often challenging to interpret in their decision-making processes. Moreover, there has been a growing apprehension regarding the presence of bias and lack of transperancy in artificial intelligence (AI) systems, as it possesses the capability to reinforce existing societal disparities. This issue has garnered significant attention and is considered a significant area of concern. Ensuring the equitable and impartial decision-making of AI systems presents a critical and urgent challenge that requires not only technical proficiency but also a comprehensive comprehension of social dynamics and ethical principles.

Artificial intelligence (AI) exerts a substantial influence on various aspects of society, including labor markets, privacy norms, and interpersonal interactions. The deployment of AI technologies has sparked discussions regarding job displacement and the changing nature of work, as automation increasingly takes over certain tasks, potentially resulting in shifts in employment patterns and skill requirements. Moreover, the acquisition and examination of substantial quantities of individualized data have elicited apprehensions regarding encroachment upon personal privacy and the safeguarding of confidential information. Striking a balance between harnessing the advantages of AI-driven insights and safeguarding individual‘s privacy rights poses a formidable challenge.

The legal domain is facing challenges in effectively adapting to the swift progressions in artificial intelligence technology. In order to tackle the issues pertaining to liability attribution in instances of AI errors or accidents, intellectual property rights concerning AI-generated content, and the possibility of AI systems engaging in autonomous decision-making with limited human intervention, there is a need for innovative legal frameworks. The convergence of artificial intelligence (AI) with established legal frameworks, regulations, and ethical principles presents intricate legal challenges that necessitate the involvement of lawmakers, technologists, and ethicists. According to a research done by microsoft, it is stated that how we design human-AI interaction is key to complementarity and empowering human agency. We need to carefully plan how people will interact with AI systems that are stochastic in nature and present inherently different challenges than deterministic systems. Designing and testing human interaction with AI systems as early as possible in the development process, even before teams invest in engineering, can help avoid costly failures and redesign [63].

The research is structured into three main sections, namely an examination of the perceived ethical,social and legal challenges associated with artificial intelligence (AI), followed by the source of these concerns including the attacks that are performed on the machine learning algorithms, and finally an exploration of potential strategies and guidelines for addressing and alleviating these concerns. The objective of this thesis is to comprehensively examine the various dimensions of ethical, social, and legal dilemmas arising in the era of artificial intelligence and gain a comprehensive understanding of the multifaceted implications of AI on society through a rigorous analysis of empirical case studies, theoretical frameworks, and existing methodologies. Through direct engagement with these concerns, the thesis aims to make a meaningful contribution towards a more holistic comprehension of the ethical, social, and legal factors that form the foundation of responsible artificial intelligence (AI) development and implementation.

[4]

2 Background and Related Work

The Age of AI signifies a pivotal juncture in human history, as it is widely believed that intelligent machines and algorithms have the potential to enhance human capabilities. This augmentation enables individuals to effectively address and resolve complex problems, thereby enriching the decision-making processes.

Significant advancements in artificial intelligence (AI) technology have been observed in the domains of natural language processing, computer vision, machine learning, and robotics. These advancements have paved the way for the utilization of AI in various sectors including healthcare, finance, transportation, and other relevant fields. The emerging innovations possess the capacity to yield advantageous outcomes for the diverse sectors discussed. Concurrently, artificial intelligence (AI) has the potential to be misused or exhibit unanticipated and potentially detrimental behaviors. Consequently, inquiries regarding the significance of law, ethics, and technology in the regulation of AI systems have become increasingly crucial. As stated by Elon Musk, CEO of Tesla, SpaceX and co-founder of OpenAI and The Boring Company, “If you’re not concerned about AI safety, you should be. Vastly more risk than North Korea.” [92].

AI can be divided into two types: weak AI and strong AI. Artificial intelligence that shows only very specific intelligence is related to what we know as «weak AI», in contrast with «strong AI», which, The philosopher John Searle was the first to introduce the distinction between weak and strong AI in a paper published in 1980 that criticised artificial intelligence (Searle, 1980) that raised, and still raises, many doubts [61]. Strong AI would imply a properly programmed computer that does not emulate a mind, but rather «is a mind». We should be clear that general AI and strong AI are not the same thing. There is a connection, of course, but only in one direction: that is, any strong AI must necessarily be general, but general AIs that are not strong can exist, meaning that they simulate the ability to show general intelligence without being real minds as interprated by López de Mántaras [61].

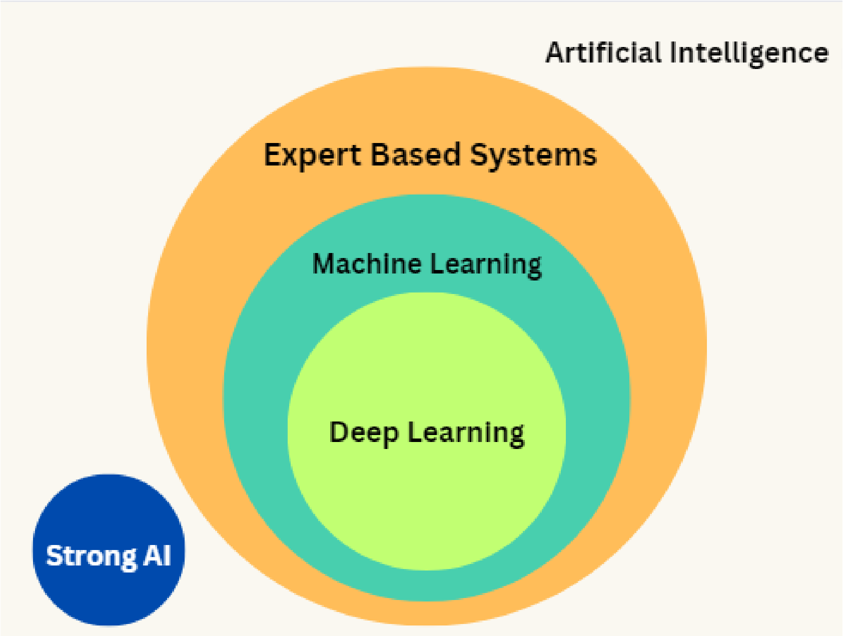


Figure 1 AI's Onion Diagram

In contrast to weak Researchers from diverse disciplines are collaboratively engaged in the development of robust artificial general intelligence (AGI), which possesses the ability to execute a wide range of tasks with human-level intelligence. This stands in contrast to narrow AI, which is limited to performing specific tasks. The topic of General AI has generated significant controversy and has engendered intense discussions among researchers due to their apprehensions regarding its potential to result in superintelligence according to Müller and Bostrom [1]. Super intelligence, as loosely anticipated by Bostrom, refers to an intellect that surpasses human cognitive abilities in nearly all areas of interest. The concept of self-improving general-purpose systems refers to the ability of a system to enhance its own capabilities and adapt to a Artificial Intelligence (AI), specifically Artificial General Intelligence (AGI), which refers to a hypothetical computational system capable of acquiring knowledge and executing a comprehensive spectrum of tasks equivalent to those performed by human beings. Through the acquisition of learning capabilities and the iterative enhancement of its own programming, an artificial intelligence system could enhance its ability to self-improve. Consequently, it could potentially acquire the capacity to circumvent any limitations imposed by its initial code and autonomously develop its own objectives. Alternatively, humans could initially equip the AI system with this capacity. For several decades, scholars and authors have extensively explored the concept of a conscious, intelligent, and purpose-driven machine with the capacity to execute a comprehensive array of tasks equivalent to those performed by humans. Irrespective of the state of consciousness or intentionality. The potential ramifications for humanity would be significant if a general-purpose machine possessing advanced intelligence and capabilities for self-improvement and self-learning were to emerge.

On the other hand, the relationship between artificial intelligence (AI) and big data is mutually influential. As stated by Pat Gelsinger, CEO of Vmware regarding the importance of big data *“Data is the new science. Big data holds the answers.“* [2]. Although existing big data analytics processes are currently implemented, the full potential of big data can only be fully harnessed by employing artificial intelligence (AI) techniques. On the contrary, big data offers artificial intelligence (AI) a substantial and heterogeneous collection of input data for the purpose of development and learning. AI and big data are closely interconnected in this context. The concept of big data lacks a universally agreed-upon definition, yet it is frequently employed to denote vast quantities of data generated and gathered in diverse formats. The term 'big data' encompasses a wide range of information, spanning various types and magnitudes, big data is the fuel that powers the evolution of AI’s decision making. Big data can be explored and analyzed for information and insights. Big data analytics is the use of processes and technologies, including AI and machine learning, to combine and analyze massive datasets with the goal of identifying patterns and developing actionable insights. This helps you make faster, better, data-driven decisions that can increase efficiency, revenue and profits.

Artificial intelligence (AI) has the capability to provide support to users throughout all stages of the big data cycle. This encompasses the activities associated with collecting, storing, and accessing a wide range of data from different origins. Artificial intelligence also has the capability to discern various data types, establish potential linkages between datasets, and comprehend information through the application of natural language processing. The tool has the capability to automate and expedite various data preparation tasks, such as generating data models, and can aid in the process of data exploration. The system possesses the ability to acquire knowledge of prevalent patterns of human errors, thereby enabling it to identify and rectify potential inaccuracies in the provided information. The system has the ability to acquire knowledge through observation of the user's interactions with an analytics program, thereby facilitating the rapid extraction of unforeseen insights from extensive datasets. Furthermore, Artificial intelligence (AI) has the capability to acquire knowledge regarding subtle distinctions in meaning or context-specific nuances. This ability enables AI to assist users in gaining a deeper comprehension of numeric data sources. The system possesses the capability to notify users of irregularities or unforeseen trends within data. It actively observes events and detects potential risks by analyzing system logs or social networking data, among other sources.

It is worth emphasizing that nearly every action performed by individuals contributes to the generation of data. This includes activities such as online searching, sharing and transmitting everyday information with government entities, companies, and social media platforms, as well as the mere act of carrying a smartphone. Consequently, substantial volumes of data, whether intentionally or unintentionally, are produced, providing extensive insights into individuals. The proliferation of the Internet of Things (IoT) is leading to the integration of network infrastructure into various aspects of our physical surroundings and individual domains. Consequently, the volume of data generated, gathered, and utilized by artificial intelligence (AI) systems will extend to encompass our personal experiences.

While it is impossible to predict the effects and outcome of Artificial General Intelligence (AGI) with absolute certainty, various scenarios can be envisaged. These scenarios encompass situations in which Artificial General Intelligence (AGI), despite its heightened intelligence and capabilities, remains subject to human oversight and is employed for the betterment of humanity. On the other hand, it is conceivable that AGI could function autonomously from human intervention and coexist harmoniously with humans. From a logical standpoint, it is plausible that there exist certain circumstances in which Artificial General Intelligence (AGI) could pose a potential risk to humanity. This risk could potentially extend to an existential level, either through deliberate or unintentional means, such as through direct or indirect harm inflicted upon humans, through acts of aggression or domination, or through the disruption of vital systems and depletion of resources upon which we rely.

One of the main issues related to AI is the issue of biasas it remains prevalent even within the realm of supervised learning, wherein the process of labeling data is frequently accompanied by pre-filtering or selection procedures. Contrary to its claim of objectivity, such a selection is susceptible to reflecting the subjective interpretation of the data curator, potentially leading to an over-representation of certain categories at the expense of others. Consequently, this can perpetuate biases related to gender, race, and other minority groups.

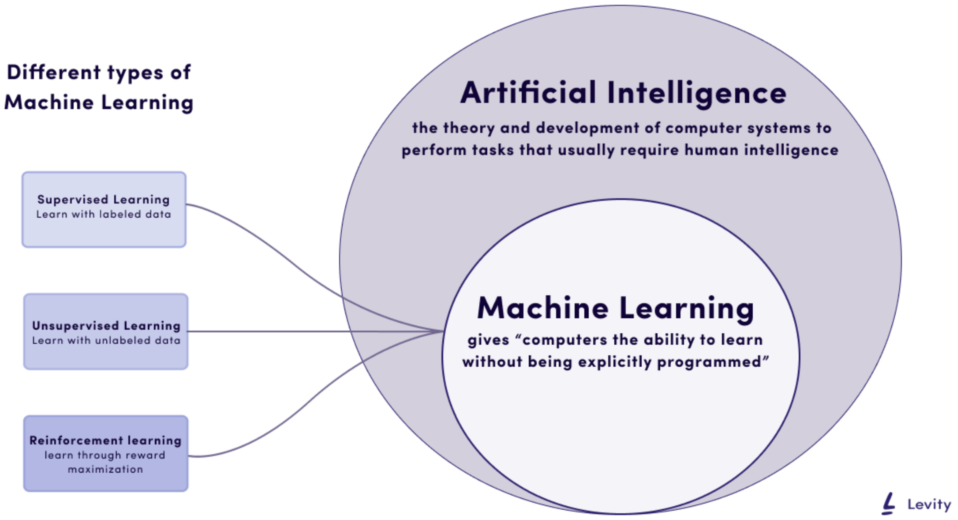

Different types of Machine Learning
Supervised Learning
Learn with labeled data
Unsupervised Learning
Learn with unlabeled data
Reinforcement learning
learn through reward maximization
Artificial Intelligence
the theory and development of computer systems to perform tasks that usually require human intelligence
Machine Learning
gives "computers the ability to learn without being explicitly programmed"
Levity

Figure 2 Different Types of ML [58]

To come to an understanding to the issues caused by AI, a careful consideration needs to be paid to the structure of machine learning. Machine learning can be classified into three main components, with one of them being supervised learning. Supervised learning is distinguished by its exploitation of labeled datasets as a core approach in the field of machine learning. The primary objective of these datasets is to streamline the process of training and overseeing algorithms in effectively categorizing data or predicting future results. Through the utilization of annotated inputs and their related outputs, the model exhibits the ability to assess its precision and improve its performance through iterative learning. The dataset is divided into two distinct subsets, specifically referred to as the training dataset and the test dataset. The objective of this partitioning is to effectively utilize the training dataset for model training and afterwards utilize the test dataset to assess the performance of the trained model. The model derived via supervised learning is later utilized to generate predictions on unannotated data that belongs to the same class as the data used for training.

Unsupervised learning employs machine learning techniques to analyze and cluster datasets that lack explicit labeling. The algorithms discussed above demonstrate the capability to autonomously identify underlying patterns within datasets, without requiring human intervention. Reinforcement learning (RL) is a specialized domain within the study of machine learning that focuses on the decision-making abilities exhibited by intelligent agents as they interact with a specific environment. The main objective of reinforcement learning (RL) is to optimize the total accumulated reward acquired by the agents involved. Reinforcement learning and supervised learning possess discernible traits that set them apart from each other. Supervised learning involves the utilization of training data that is accompanied by an answer key, which facilitates the training of the model by providing the correct answers. In contrast, reinforcement learning does not possess a clearly defined answer key, as the reinforcement agent independently determines the behaviors required to successfully complete a specified task. In instances where a training dataset is not accessible, the system is compelled to acquire information through its own experiential learning.

Looking into reinforcement learning, the intelligent agent, commonly referred to as an IA, is an entity that exhibits intelligent behavior. This agent is capable of perceiving its surrounding environment, making independent decisions and taking actions to accomplish specific objectives. Furthermore, it has the ability to enhance its performance through learning or the acquisition of new knowledge. The establishment of human trust in the tools we utilize is contingent upon comprehending the processes involved in the tool's creation and/or the processes that establish its functionality. The agents engage in self-training by utilizing reward and punishment mechanisms. The notion refers to the strategic decision-making process aimed at maximizing favorable results and minimizing unfavorable repercussions, through attentive observations within a specific context. It serves as a catalyst for both favorable and unfavorable behaviors. Essentially, an agent, or perhaps multiple agents, is designed with the ability to perceive and understand the immediate environment in which it is located. Moreover, it demonstrates the capacity to perform tasks and participate in various activities within this particular context. The acquisition of task execution abilities by an intelligent agent is achieved through iterative and experimental interactions with a dynamic environment. This specific learning process enables the agent to independently develop a series of decisions that maximize a reward metric for a specified objective, without any human interaction or explicit programming to complete the task.

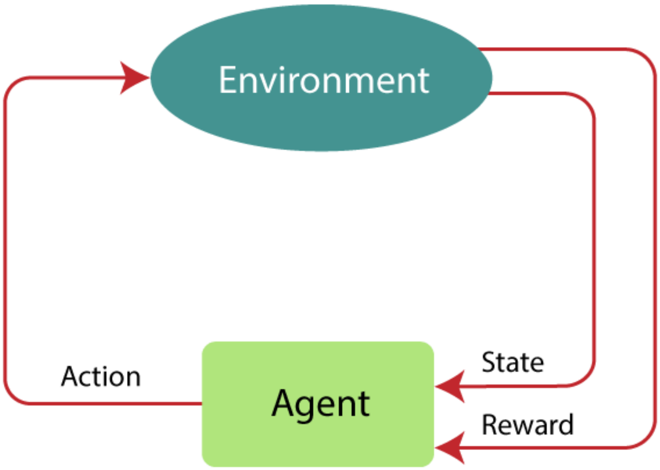


Figure 3 Reinforcement Learning [68]

2.1 The Social, Ethical and Legal Challenges Associated With Artificial Intelligence (AI)

2.1.1 The Ethical Challenges Associated With AI

as referred to by Siau and Wang at the moment, AI is often referred to as Narrow AI or Weak AI [8]. It is effective in a restricted and particular domain. The performance of narrow AI is heavily reliant on training data and programming, both of which are closely related to big data and individuals. The building of AI systems is heavily reliant on massive amounts of data, including private as well as personal data. These data must be properly safeguarded to avoid improper use and abuse. AI, particularly machine learning and deep learning, is not always transparent to scrutiny. AI may evolve without human monitoring and guidance due to the black box that humans are unable to interpret. The lack of transparency also raises the possibility of malicious use.

The most important factor is human bias, such as gender bias and race bias according to Koolen and Cranenburgh [93], which AI may inherit. Because

AI systems are still being trained by humans and using human-created datasets, existing biases may be learned by AI systems and displayed in real-world applications.

Another issue to consider is accountability. Who should be held accountable when an AI system fails to complete a specific task? This could result in what is known as *"the problem of many hands"* which was brought up by Timmermans et al. [94]. The utilization of an artificial intelligence system can potentially yield an undesirable effect, which can be attributed to several factors such as flaws in the programming code, inaccuracies in the inputted data, wrong execution of operations, or other relevant variables. Which party bears responsibility for the adverse outcome: the programmer, the data owner, or the end users? The escalating utilization of artificial intelligence (AI) in contexts with significant risks has engendered a heightened demand for AI design and governance that is characterized by accountability, equity, and transparency. The two primary dimensions of transparency encompass the accessibility of information and its comprehensibility. The accessibility of information regarding algorithm functionality is often deliberately obscured. The utilization of artificial intelligence (AI) has the potential to result in a lack of identifiable individuals who can be held accountable for any resulting harm or negative consequences. The extent of the threat is uncertain, and the utilization of machines will significantly impede our capacity to attribute accountability and assume responsibility for decision-making.

On the other hand, the problem of selection bias in datasets used to build AI algorithms is common. Buolamwini and Gebru [70] established that there is bias in automated facial recognition and the associated datasets, resulting in lower accuracy in recognizing darker-skinned individuals, particularly women. A large number of data points are required for ML, and the majority of clinical trial research databases are drawn from specific populations. As a result, when applied to underserved and thus likely underrepresented patient groups, the resulting algorithms may fail more frequently. In 2016, Microsoft's Tay chatbot exhibited racist behavior subsequent to being trained by Twitter users. This inquiry pertains to the determination of ethical norms that businesses ought to adhere to in the context of building artificial intelligence (AI) systems.

When it comes to the research field, one main area that has been gaining attention as the time passes is the prediction of an industry. AI is currently based on various algorithms and techniques such as supervised learning, unsupervised learning, and deep learning. These various approaches result in AI-based systems with varying self-learning velocities. Differences in the quality and quantity of training data also mean that the capabilities of AI in organizations vary greatly, emphasizing why we believe that the ability to self-learn is a key distinguishing feature of AI in organizations.

In relation to the ethical governance of artificial intelligence (AI), the emergence of a classification system based on the rate of autonomous learning presents a quandary regarding the systematic incorporation of ethical considerations within organizational contexts. The increasing proximity of human-AI interactions is widely acknowledged to heighten the significance of ethical problems. AI techniques that enhance the effectiveness of analyzing extensive datasets, including recommendation, forecasting, and optimization algorithms, are not accorded significant ethical attention as their direct impact on human beings is deemed negligible. Furthermore, artificial intelligence possesses the capability to engage in direct communication with users, as exemplified by the utilization of chatbots in customer service. Additionally, it has the potential to exert influence on the lives of individuals and minority groups, potentially resulting in adverse consequences such as the unfair disadvantage of job seekers during the interview process within the field of human resources or the manipulation of self-driving automobiles.

Transparency is also one of the primary concerns raised by artificial intelligence. One of the greatest obstacles posed by artificial intelligence (AI) systems is the lack of transparency surrounding their operational mechanisms and decision-making processes. The absence of transparency presents a significant obstacle to identifying potential biases or ethical concerns that may arise in the implementation of AI systems. Businesses must demonstrate the operational and decision-making transparency of their artificial intelligence (AI) systems in order to effectively address this challenge. A plausible strategy involves providing detailed explanations of the system's decision-making methodology. This strategy has the potential to increase the level of trust among customers and other stakeholders who may be hesitant to embrace or rely on AI systems due to a lack of understanding.

One of the primary sources of ethical dilemmas stems from the concept of the "black box," which is derived from the idea that artificial intelligence systems and machine learning models function in a manner that is hidden from human comprehension, resembling the contents of an opaque and sealed container. These systems are constructed utilizing elaborate mathematical models and extensive datasets encompassing numerous dimensions, resulting in the creation of deep linkages and patterns that regulate their decision-making mechanisms. Nevertheless, individuals face a lack of instant access to or comprehension of these internal mechanisms. The AI black box dilemma poses distinct issues when examined from a regulatory perspective. The challenge encountered by regulators in evaluating the adherence of artificial intelligence (AI) systems to existing norms and guidelines arises from the restricted transparency in their functioning. Moreover, the absence of transparency can hinder regulators in their efforts to establish progressive frameworks capable of effectively addressing the dangers and problems presented by artificial intelligence (AI) applications.

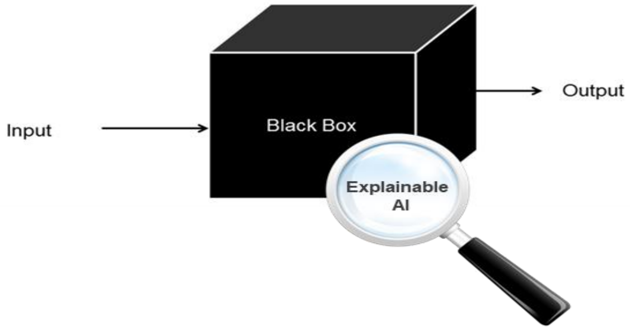


Figure 4 BlackBox Problem [77]

The AI black box problem refers to the challenge of comprehending the underlying rationale behind the predictions or choices produced by an AI system in real-world contexts. The aforementioned issue is commonly found in deep learning models, particularly neural networks, wherein numerous layers of interconnected nodes are involved in the hierarchical processing and transformation of data. The complex structure of these models and the nonlinear changes they execute provide significant difficulties in comprehending the fundamental reasoning behind their outputs.

The assignment of accountability and responsibility is a fundamental problem because it is difficult to determine fault when errors or biases occur because Blackbox AI is opaque as interpreted by Floridi & Cowls [71]. As the internal workings of these systems may not be discernible to human evaluators, hidden biases and the resulting unfair or discriminatory outcomes represent another urgent concern [72]. Additionally, Blackbox AI's lack of interpretability and explain-ability impedes transparency, making it difficult to comprehend how decisions are made [73]. When AI affects people's lives, obtaining informed consent becomes challenging, and the inability to disclose decision-making criteria raises ethical concerns [74]. The ethical issues with Blackbox AI are further exacerbated by discrimination based on unspoken criteria, worries about automation bias, and the complexity of regulatory compliance are rising [75][76][77]. In order to overcome these obstacles, transparency, fairness, and interpretability in AI systems must be improved. This is a continuous goal of research and policy initiatives that support the creation and use of ethical AI.

Highlighting the ethical management of AI, one of the studies that has been done by Brendel, Mirbabaie, Lembcke, and Hofeditz [69], as interpreted, an AI positioning matrix has been proposed that integrates two perspectives:
1. The degree and velocity of self-learning.

2. The degree of AI’s impact on humans.

The cases were chosen and classified as examples to describe the range of current AI-based technologies; they do not claim to be complete, nor are they based on concrete numbers. The first sector includes AI-based technologies that is believed to have little self-learning and have little impact on humans. Because of the lower level of human-AI closeness and self-learning, the likelihood of AI-caused impactful errors is also lower. As a result, this sector is not considered to be particularly important for the ethical management of AI. The second sector includes all cases with a medium level of self-learning and human impact. Because these technologies have a greater impact on people's lives and behavior, this sector is more relevant for the ethical management of AI. The third sector includes AI technologies that have a significant impact on people's lives and have a high degree of self-learning. Therefore, by employing the positioning matrix outlined earlier, one can evaluate and prioritize different AI-related initiatives based on their potential impact on human beings and society, as well as the necessity of conducting ethical assessments for these undertakings. The adherence to ethical standards is crucial for artificial intelligence (AI), particularly when considering its substantial capacity to impact persons and societies, as well as its advanced capabilities for self-learning.

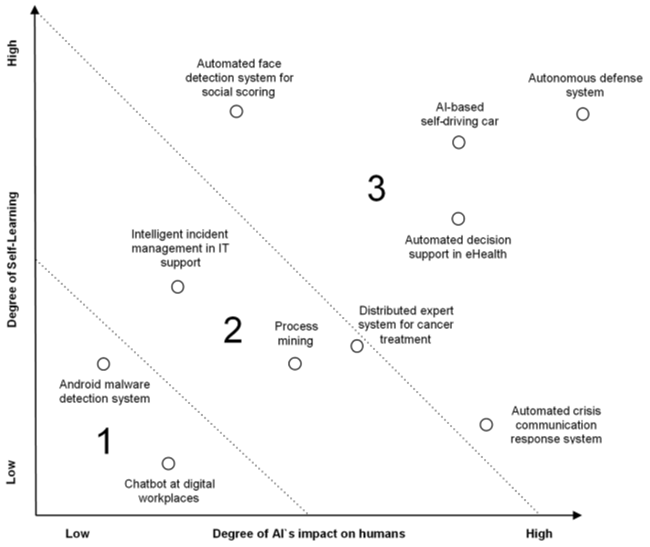

High
Degree of Self-Learning
Low
Automated face detection system for social scoring
AI-based self-driving car
Autonomous defense system
3
Automated decision support in eHealth
Intelligent incident management in IT support
2
Process mining
Distributed expert system for cancer treatment
Android malware detection system
1
Automated crisis communication response system
Chatbot at digital workplaces
Low
Degree of AI`s impact on humans
High

Figure 5 AI Ethics Matrix [6]

Deep Fake AI

One of the primary obstacles encountered in the transition from curation to creation pertains to the mitigation of biases and the establishment of fairness in AI-generated content. Artificial intelligence (AI) systems acquire knowledge from extensive datasets, and if these datasets contain biases, those biases can be replicated in the content generated by the AI systems, leading to the creation of "real time" deepfakes. The emergence of deepfakes, which are powered by artificial intelligence (AI), has elicited apprehension due to their capacity to manipulate and exert influence on electoral processes. These sophisticated fabrications possess the ability to generate highly convincing yet falsified content, thereby misleading voters and undermining the democratic framework. However, the concerns surrounding deepfake technology extend beyond its implications on electoral processes. Additionally, this phenomenon gives rise to significant apprehensions regarding the potential utilization of manipulated media in the context of warfare. In such scenarios, the dissemination of false information and the presentation of fabricated evidence have the potential to intensify existing tensions, escalate conflicts, and yield profound and wide-ranging ramifications.

Deepfake technology, renowned for its ability to deceive, derives its nomenclature from deep learning, a variant of artificial intelligence (AI). Deepfake artificial intelligence (AI) employs deep learning algorithms, which possess the ability to autonomously acquire problem-solving skills through extensive datasets. These algorithms are utilized to replace faces in videos, images, and various forms of digital content, thereby creating a counterfeit representation that closely resembles reality. Based on the interpretation derived from a survey that has been conducted by multiple researchers regarding deepfakes, In order to perform face swapping between source images and target images, it is necessary to utilize two sets of encoder-decoder pairs. Each pair is trained on a specific image set, and the encoder's parameters are shared between the two network pairs. To clarify, both pairs possess an identical encoder network [52]. The content is generated by employing two algorithms that are in competition with each other. One of the components is commonly denoted as a generator, whereas the other component is commonly denoted as a discriminator. The generator is accountable for generating artificial digital content, whilst the discriminator is assigned with the responsibility of distinguishing between genuine and simulated content. Once the discriminator effectively differentiates between genuine and altered information, it shares this acquired knowledge with the generator to improve future generations of deepfakes.

When combined, these two algorithms constitute a generative adversarial network known as GAN. The system employs a series of algorithms to undergo training in order to develop the ability to identify patterns, thereby facilitating the acquisition of essential attributes required for generating synthetic images [53].

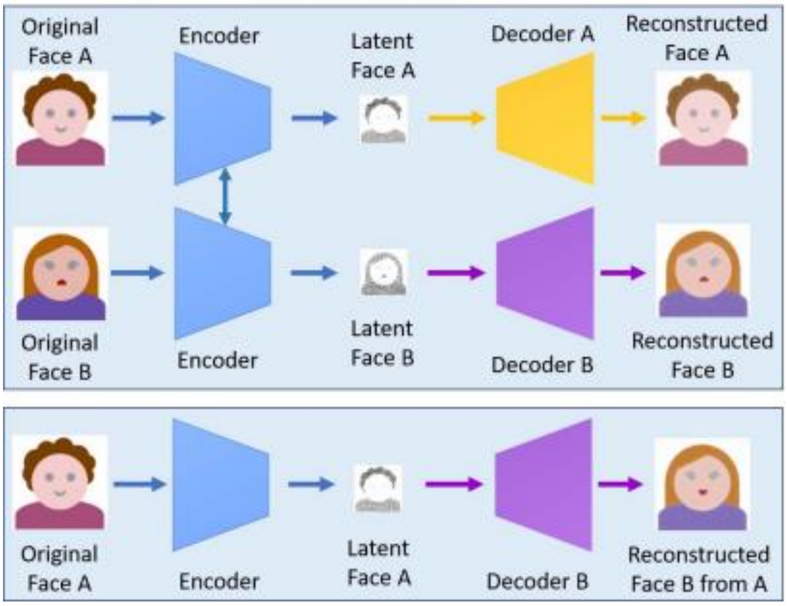


Figure 6 Deep Fakes [52]

Deepfakes have the potential to be exploited in order to incite political or religious tensions between nations, deceive the general public, manipulate election outcomes and campaign results, or induce disorder in financial markets through the dissemination of fabricated news. The technique possesses the capacity to produce synthetic satellite imagery of the Earth that portrays fictitious entities, aiming to induce perplexity among military experts. For example, it can be utilized to generate a simulated bridge that extends across a river, even in the lack of a physical bridge in the real world. The progress made in deep neural networks, along with the vast availability of data, has led to the creation of manipulated photos and movies that possess a high degree of realism, making them extremely difficult to distinguish from authentic content for both people and advanced computer algorithms. As interpreted, the generation of manipulated images and videos has become significantly more accessible in contemporary times, requiring only a minimal input such as an identity photo or a brief video featuring the targeted individual [52].

2.1.2 The Social and Legal Challenges of AI

The advent of AI has marked a profound technological revolution, characterized by the ability of machines to mimic human cognitive functions, such as learning, problem-solving, and decision-making. From virtual assistants to autonomous vehicles, AI systems have showcased their potential to transform industries, streamline operations, and revolutionize the way we interact with technology. As AI technologies permeate diverse sectors, the legal landscape faces the pressing challenge of providing an appropriate framework for governing these transformative systems. Addressing issues of liability, responsibility, and accountability becomes critical when AI-powered decisions impact human lives and societal outcomes. Striking a balance between promoting innovation and safeguarding individual rights and public interest is a complex task that necessitates multidisciplinary collaboration and forward-thinking policies.

The rapidly evolving nature of AI presents a unique challenge for regulatory bodies worldwide. Existing legal frameworks may struggle to keep pace with the rapidly changing technology, raising questions about the adequacy of current regulations in addressing AI-related issues. As AI systems become increasingly autonomous and independent, there is a growing need to establish clear and robust guidelines for AI developers, users, and stakeholders to ensure ethical and responsible AI deployment.

Beyond legal concerns, the widespread adoption of AI has profound social implications that require careful examination. The automation of tasks previously performed by humans can lead to job displacement, necessitating reskilling and adaptation of the workforce. Additionally, the potential exacerbation of societal inequalities due to biased algorithms and data-driven decision-making warrants scrutiny to ensure equitable outcomes. AI's profound societal implications extend beyond the legal realm. The widespread adoption of AI technologies has the potential to reshape the job market, leading to concerns about job displacement and the need for upskilling and reskilling the workforce. Furthermore, the pervasive influence of AI on decision-making processes raises questions about fairness and social justice, as algorithms may perpetuate biases and exacerbate existing inequalities.

Data privacy and protection

Data is all around us and is generated by almost everything we do. The first is data that we voluntarily share, and the second is data that is generated literally every time we do something - whether we travel, order a meal, or use transportation. Without a doubt, this data is extremely valuable, and several companies are willing to pay for access to it. The use of Artificial Intelligence (AI) in everyday life raises a number of data protection concerns. All such concerns must be carefully considered in light of the provisions of EU Regulation 2016/679, General Data Protection Regulation (GDPR). The purposes of personal data processing and the enhancement of AI systems involve a material issue in terms of data processing purposes.

In 2017, Genpact surveyed over 5,000 people from various countries and discovered that 63% of respondents valued privacy over a positive customer experience and wanted businesses to avoid using AI in case it invaded their privacy [15]. The main concerns here are how an AI system can access a consumer's personal information, what kind of information it can access, and how serious a privacy violation this could be.

Invasive surveillance, which can erode individual autonomy and exacerbate power imbalances, and unauthorised data collection, which can compromise sensitive personal information and leave individuals vulnerable to cyber-attacks, are two of these issues. These issues are frequently exacerbated by the power of BigTech companies, which have access to vast amounts of data and significant influence over how that data is collected, analysed, and used. A considerable proportion of the value of artificial intelligence (AI) stems from its ability to discern patterns that elude human perception, acquire knowledge about individuals and collectives, and generate predictions. Artificial intelligence (AI) has the capability to generate data that would otherwise be challenging to collect or non-existent. Consequently, the data that is collected and utilized may extend beyond the explicit information that an individual has intentionally revealed. The potential of predictive technologies lies in their capacity to derive meaningful insights from seemingly disparate and innocuous data points. As an illustration, an artificial intelligence (AI) system designed to enhance the efficacy of the recruitment procedure could potentially deduce an applicant's political inclinations based on supplementary data and subsequently factor this information into the decision-making process.

The issue of information inference gives rise to concerns regarding the precise definition of personal information, as well as the appropriateness of assuming personal information about individuals who have opted not to disclose it. Additional concerns are raised regarding the ownership of the aforementioned data and its adherence to information privacy principles, including the requirement to notify individuals when data pertaining to them has been collected through inference.

AI techniques, specifically machine learning, heavily depend on extensive datasets for the purpose of training and evaluating algorithms, as per their inherent nature. The accumulation of substantial quantities of data has the potential to facilitate advancements in artificial intelligence (AI) development. However, it may also directly conflict with the principle of limiting data collection. The collection of data for AI systems is often not obtained through conventional transactions where individuals consciously disclose their personal information to a requesting party, due to the technological progress in IoT devices, smartphones, and web tracking. Indeed, a considerable number of individuals frequently lack awareness regarding the extent of data gathered from their devices, which subsequently serves as input data for artificial intelligence (AI) systems.

The extent to which the restriction of personal information collection aligns with the operational capabilities of AI technologies and data-collecting devices is constrained. Nevertheless, the vast accumulation of such data inherently presents inherent privacy concerns. Artificial intelligence (AI) has the potential to enhance individual’s capacity to articulate their preferences regarding the utilization of their personal information. For instance, it is conceivable to imagine the existence of services that possess the capability to acquire knowledge of their user’s privacy preferences and subsequently implement distinct conditions pertaining to the collection of data concerning various individuals. AI has the potential to significantly contribute to the creation of personalized models based on individual preferences. These models can effectively address the objectives of information privacy law, including transparency, consent, and reasonable expectations. In fact, AI may surpass the current notice and consent model in terms of effectiveness. The use limitation principle ensures that personal information is only used for the purpose for which it was collected once it has been collected. In general, organizations may use personal information for a secondary purpose that is reasonably expected' by the individual. This begs the question of whether information used as input data for an AI system can be considered a reasonably expected secondary purpose,' given that the outcome is often unknown to the individual. Just as AI can reveal patterns and relationships in data that humans are unaware of, it can also reveal new potential uses for that information.

Threats of AI to human existence:

The main set of issues arises from the capacity of artificial intelligence (AI) to efficiently analyze, categorize, and evaluate extensive datasets that comprise personal information, such as photographs obtained through the growing prevalence of cameras. This functionality enables the development of customized and focused marketing and informational initiatives, together with the deployment of advanced surveillance systems. This ability of AI can be put to good use, for example, improving our access to information or countering acts of terrorism as stated by multiple researchers from BKJ Global Health [78]. But it can also be misused with grave consequences. The use of this power to generate commercial revenue for social media platforms, for example, has contributed to the rise in polarisation and extremist views observed in many parts of the world [79].

It has also been harnessed by other commercial actors to create a vast and powerful personalised marketing infrastructure capable of manipulating consumer behaviour. Experimental evidence as stated by (Agudo and Matute, 2021) has shown how AI used at scale on social media platforms provides a potent tool for political candidates to manipulate their way into power [80]. The convergence of deepfakes, a technology capable of manipulating or distorting reality, with AI-driven information systems has the potential to further amplify the erosion of democratic norms. This phenomenon may manifest as a pervasive breakdown of trust, alongside the propagation of social fragmentation and strife, ultimately resulting in detrimental implications for public health. The possible implementation of AI-powered surveillance systems may be utilized by governmental institutions and other prominent groups as a mechanism to exercise increased control and tyranny over humans. An illustrative instance that serves as a manifestation of this phenomena is the implementation of the Social Credit System in China. The proposed method combines sophisticated facial recognition software with the examination of comprehensive datasets encompassing people' financial activities, movements, police histories, and social relationships. By means of this extensive assessment, the system produces evaluations of individual behavior and reliability, later resulting in the automatic enforcement of penalties on those who are determined to have displayed unacceptable behavior.

This particular AI application has the potential to further amplify social and health disparities, thereby perpetuating individual’s confinement within their current socioeconomic positions.The development of autonomous weapons is currently progressing at a rapid rate. Prominent nations are engaged in a competitive struggle to establish dominance in the realm of autonomous weaponry. Various forms of automated warfare technology, such as advanced jets and robotic soldiers, have already been developed and are currently deployed in the field. The technological advancements have significantly augmented the capacity for destruction, surpassing quantifiable limits. The country's national security is focused on priority areas such as nuclear, aerospace, cybersecurity, and biotechnology. For instance, the significance of cybersecurity is increasing, as the scope of the conflict extends beyond traditional battlefields to encompass the systems we construct. Major nations such as the United States, China, and Russia are engaged in a competitive pursuit of global supremacy in the field of artificial intelligence (AI) technology [54].

The following series of inquiries relates to the progress of Lethal Autonomous Weapon Systems (LAWS) [78]. The incorporation of artificial intelligence (AI) within military and defense systems encompasses a diverse array of applications, several of which possess the capacity to augment security measures and promote peaceful resolutions. Nevertheless, the potential benefits of LAWS are eclipsed by the risks and threats they present. According to research done by BMJ, weapons have the ability to autonomously detect, select, and engage human targets without requiring human supervision [81]. The concept of dehumanizing lethal force is widely recognized as the third significant transformation in the domain of warfare, following the initial adoption of gunpowder and the subsequent advancement of nuclear weapons. Lethal autonomous weapons manifest diverse dimensions and arrangements. Nevertheless, it is crucial to acknowledge that these devices are equipped with armaments and explosives that can be attached to small and agile contraptions like quadcopter drones. The drones possess sophisticated cognitive abilities, allowing them to independently operate and accurately perceive and navigate their environment. Moreover, it is important to highlight that these weapons exhibit the capacity for efficient and economical mass manufacturing, enabling their swift deployment to inflict extensive casualties on a large scale [82]. One can conceive of a hypothetical situation in which a considerable quantity of diminutive unmanned aerial vehicles (UAVs), possessing explosive functionalities, visual identification capabilities, and self-governing navigation capabilities, could be housed within a conventional shipping container. According to reference [82], it is possible to program these unmanned aerial vehicles (UAVs) to execute acts of mass violence autonomously, without the need for human supervision. Lethal autonomous weapons systems (LAWS) exhibit resemblances to chemical, biological, and nuclear weapons, as they embody a novel form of widespread devastation. These systems possess the benefit of cost-effectiveness and the potential to accurately detect and engage specific targets. The aforementioned findings have substantial implications for the future conduct of armed conflicts, as well as for the broader aspects of international, national, and personal security. There has been a continuous discourse across multiple platforms concerning the adoption of tactics aimed at mitigating the proliferation of lethal autonomous weapon systems (LAWS). Furthermore, there have been discussions regarding the feasibility of protecting these systems from cyber intrusion, unintended triggering, and deliberate abuse.

The social determinants of health have the potential to be negatively impacted by artificial intelligence (AI), thereby affecting the well-being of a substantial portion of the population. This phenomenon may manifest itself through the exercise of authority and manipulation over individuals, the utilization of lethal autonomous weapons, and the negative psychological effects stemming from extensive joblessness in the event that AI-powered systems replace a significant portion of the workforce.

In addition to the possibility of extinction, there is a legitimate concern regarding the potential for a civilization to become indefinitely entrenched in a flawed trajectory. An exemplar case, commonly known as "value lock-in," arises when the integration of AI into society is hindered by persistent moral deficiencies reminiscent of historical periods of slavery. This scenario has the potential to perpetuate and consolidate these deficiencies, consequently impeding the progress of moral principles. The utilization of artificial intelligence (AI) holds promise for the dissemination and preservation of the value systems upheld by its creators. The evaluation of the consciousness and level of consciousness in a highly developed artificial intelligence (AI) presents considerable difficulties, making it an intricate and potentially unachievable undertaking. Nevertheless, if a substantial number of conscious machines are manufactured in the future, neglecting to prioritize their long-term welfare could conceivably result in a perilous risk to our survival.

The issue of Bias and Discrimination

Another issue raised by AI technology is the possibility of bias and discrimination. As potential biases and discrimination become more apparent, the increasing use of AI systems has intensified conversations about fairness and bias in artificial intelligence. This section examines the causes, effects, and mitigation strategies pertaining to AI's fairness and bias.

Bias occurs when a machine-learning model produces discriminatory outcomes towards particular groups or individuals. These groups are commonly characterized by a history of experiencing discrimination and marginalization based on factors such as gender, social class, sexual orientation, or race, although exceptions exist. The potential cause of this outcome may stem from flawed assumptions made during the model's development process, or from the utilization of training data that is non-representative, inaccurate, or fundamentally incorrect. It is imperative to acknowledge that bias denotes a departure from the established standard and does not invariably result in discrimination. The level of bias present in AI systems is directly influenced by the bias inherent in the training data. Consequently, if the training data exhibits bias, the resulting AI system will also exhibit bias. This phenomenon may lead to the formulation of biased judgments influenced by variables such as race, gender, or socioeconomic standing. In order to mitigate bias, it is imperative that AI systems undergo training using a diverse array of data and undergo regular audits.

One of the foremost ethical considerations regarding artificial intelligence (AI) pertains to its capacity to engage in discriminatory practices, perpetuate biases, and further amplify preexisting disparities. Due to the reliance on pre-existing data for training, algorithms have the potential to inadvertently perpetuate undesirable patterns of unfairness, which can be attributed to the information they have assimilated. The various phases involved in the machine learning pipeline, including data collection, algorithmic development, and user interaction, can potentially give rise to biases in artificial intelligence.

The three primary sources of bias commonly observed in artificial intelligence are data bias, user bias, and algorithm bias. Data bias refers to the presence of inadequate or unrepresentative data used for training machine learning models, leading to the production of biased outputs. This phenomenon may arise if the data were collected from sources exhibiting bias, or if the data were incomplete, lacking essential information, or containing inaccuracies. The ultimate outcome of a learning system is contingent upon the data it is provided with [16]. It is widely held that the substantial quantity of occurrences will surpass any inherent human prejudice. However, in the event that the training set exhibits a skewed distribution, the resulting outcome will also display a corresponding skewness. In recent times, researchers have observed the presence of bias in image recognition systems that are based on deep learning algorithms. The misinterpretation of Asian facial features by Nikon [83] as an example, encountering difficulties with skin tone can be attributed to the utilization of biased training datasets. Although both errors are fixable and unintentional, they serve as examples of the challenges that can arise when we disregard the biases present in our data.

Algorithmic bias refers to the occurrence of biases in machine learning models that arise from the utilization of algorithms with inherent biases, resulting in biased outputs. This issue may occur when algorithms are based on biased assumptions or when they make conclusions using biased criteria. Bias can be observed in algorithmic systems as a result of pre-existing cultural, societal, or institutional norms, technological constraints in their design, or unforeseen usage scenarios or user groups that were not sufficiently taken into account during the software's initial creation. The concept of "algorithmic prejudice" refers to the recurring and systematic flaws that result in unjust consequences, such as the biased favouritism towards a particular arbitrary user group in comparison to others [9]. For instance, if the algorithm has a constant tendency to recommend loans to a particular subset of users while continuously declining loans to a similar subset of users, despite the absence of creditworthiness-related factors, it might be considered as indicative of prejudice. There are numerous methods for introducing bias into an algorithm. During the compilation of a dataset, data may be collected, digitized, adapted, and entered into a database based on cataloguing criteria designed by humans. Next, programmers assign priorities or hierarchies for how a program evaluates and sorts this information. This requires human judgment regarding the categorization of data, as well as the inclusion or exclusion of certain data.

Some algorithms collect their own data based on human-selected criteria, which may also reflect the bias of their human developers. Other algorithms may reinforce stereotypes and preferences when processing and displaying "relevant" data for human users, for example, by selecting information based on prior decisions of an identical user or a collection of users. [9] For a system to be considered trustworthy, it is imperative that the output of its learning algorithms possess moral relevance. The concept of moral relevance is subject to debate, particularly when considering decisions that have legal or critical implications. The presence of algorithmic bias, resulting from the selective emphasis or avoidance of certain features, may give rise to concerns regarding its impartiality. One instance of algorithmic bias can be observed in the context of self-driving cars, where artificial intelligence (AI) models are entrusted with the critical task of ensuring human safety in the event of a collision. The potential bias arising from focus may be attributed to the design choice of prioritizing either passenger safety over pedestrian safety, or vice versa. The design choices made in the development of the algorithm have the potential to result in moral bias. AI models are frequently repurposed for various applications. The alteration of the context may lead to a departure from established statistical or moral norms.

Another form of algorithmic bias that is frequently misinterpreted as user error. The phenomenon of interpretation bias arises when there is a discrepancy between the output generated by a model and the information that is necessary for the user of the model. The model user has the potential to be either a human or another component within the system. Once again, a frequently reported bias, often arises due to the potential incompleteness of system semantics during the development phase or the contextual changes that occur over time. Consider the case of a surveillance system that employs human activity recognition. One potential approach to enhance the recognition process involves the utilization of an alternative algorithm that assesses the level of threat associated with the identified human figure. The potential for misinterpretation of detection and risk prediction within the surveillance system may lead to biased outcomes against certain individuals. The occurrence of user bias is the consciously or unconsciously introduction of a user's own biases or prejudices into an artificial intelligence system. This can occur if users provide biased training data or if their interactions with the system reflect their own biases. Human bias is a systematic way of thinking that influences the choices we make and conclusions. When a factor goes wrong, blaming external factors is a common example of this bias that affects most people. Even with good intentions, these human biases can be projected onto artificial intelligence technology. Thus, our human biases are incorporated into the technology we create in a variety of ways.

Human bias can emerge in various forms, one of which is cognitive error-induced human bias. Cognition encompasses the integration of logical reasoning and intuitive perception. While intuition can assist in making quick decisions, reason enables deliberate and intentional behaviour. The cognitive processes of individuals may exhibit a tendency to employ heuristics, leading to the formulation of hasty judgments based on initial data. The potential for bias may arise when supplementary evidence that provides support or contradiction is excluded. In this scenario, if the individual possessed prior knowledge of the high cost of their preferred dwellings, they may streamline their selection process by taking this factor into account. In addition to the phenomenon of cognitive anchoring, wherein individuals excessively rely on their initial understanding, there is also a tendency to dismiss contrary viewpoints. Relying heavily on past experiences might lead to a tendency to Favor information that confirms preexisting beliefs, so introducing bias. The inclination towards expeditious decision-making is further influenced by the presence of accessibility bias, wherein users are swayed by information that has been recently obtained or is readily available. The occurrence of human bias may be attributed to the processing of data. The cognitive ability of individuals to analyse given information is constrained. In situations where individuals are confronted with an excessive amount of information, it is likely that they will prioritize the allocation of their cognitive resources towards tasks that are seen more significant or pressing. Additionally, human bias may also arise from prejudiced attitudes. As individuals progress in their development, they tend to form assumptions that are influenced by the societal and environmental factors that surround them. Unconscious bias is influenced by established stereotypes and the corresponding level of familiarity with the underlying ideas. In the event that the user possesses a preexisting bias against the security of specific places, they would refrain from engaging with the listings, irrespective of the current state of order and security. Consequently, the model would acquire bias from the analysis of user browsing behaviour.

The presence of bias in Artificial Intelligence (AI) has significant implications for the complex ethical, social, and legal dynamics that underpin the interaction between humans and technology. The presence of bias in AI algorithms is frequently observed, and this bias can be attributed to the underlying biases inherent in the training data employed for instructing these systems. As a result, artificial intelligence (AI) technologies have the potential to sustain and magnify prevailing societal biases, resulting in discriminatory consequences within domains such as employment, lending, and law enforcement. The pernicious impact of this influence has extensive implications, as it undermines confidence in AI systems and exacerbates societal divisions. AI bias has the potential to result in inequitable treatment, the exclusion of opportunities, and the limitation of individual's potential, all of which are influenced by factors that are outside of their control. The imperative to address bias in artificial intelligence (AI) arises not only from the need for equitable technological advancement, but also from the importance of protecting human rights and preserving human dignity. By recognizing and addressing these biases, society can effectively utilize the transformative capabilities of artificial intelligence while maintaining the core values of fairness, justice, and inclusivity.

The development and deployment of artificial intelligence systems necessitate diligent attention and mitigation of diverse biases. These biases include historical biases that are based on societal disparities that are evident in data collection, representation bias that arises from insufficient diversity in the populations included in datasets, measurement bias that is influenced by human choices in selecting and analysing features, evolution bias that arises from disproportionate benchmarks, the Simpson paradox that introduces bias when analysing heterogeneous data, sampling bias that results from non-random sampling, content production bias that is tied to the characteristics of the users, and algorithmic bias that emerges from the algorithm itself rather than the input data. It is imperative to acknowledge and tackle these diverse biases in order to ensure the ethical and fair progress of artificial intelligence technologies as interpreted by Emilio Ferrara [87].

The process of data mining, one of the ways algorithms collect and analyse data , can already be discriminatory as a start, because it decides which data are of value, which is of no value and its weight—how valuable it is. The decision process tends to rely on previous data, its outcomes and the initial weight given by the programmer. One example can be when the word woman was penalised, by being given a negative or a lower weight, on a CV selection process based on the data of an industry traditionally dominated by men like the tech industry. The outcome ended discriminating women in the selection process. Some ML models, like supervised learning, learn by examining previous cases and understanding how data are labelled, which is called training. Training data that are biased can lead to discriminatory ML models. It can happen in two ways in accordance with the interpretation of Lorenzo Belenguer [16]:

1.A set of prejudicial examples from which the model learns or in the case of under-represented groups which receives an incorrect or unfair valuation.

2.The training data are non-existent or incomplete

The issue of job displacement for the workforce

With the progression of AI systems, their ability to carry out tasks that were traditionally executed by humans is steadily improving. The phenomenon may lead to the displacement of jobs, disruption of the economy, and necessitate the retraining of individuals for alternative roles. However, the issue of job loss is inextricably linked to privacy in a number of ways. For one thing, the economic disruption caused by AI technology can increase worker’s financial insecurity. As a result, individuals may be forced to give up their privacy in order to make ends meet. In the past decade, human has relied on physical work and investing their time to earn. With the technology still rising and smart intelligent robots have been developed, we may see a future where individuals are being paid just for being citizens which will be important in helping to combat the job-stealing automation [55]. AI-related jobs are now dramatically increasing in demand, according to the annual AI index report [56].

Estimations regarding the magnitude and velocity of job displacements resulting from automation driven by artificial intelligence (AI) vary significantly, spanning from tens to hundreds of millions within the next ten years. The outcome will be contingent upon the rate at which AI, robotics, and other pertinent technologies advance, alongside the policy choices enacted by governmental bodies and society at large. It is widely expected that in the current decade, the implementation of AI-driven automation will have a greater impact on low and middle-income countries compared to high-income countries [85]. This impact will primarily manifest in the replacement of jobs that require lower levels of skill. As time progresses, the automation will gradually extend its reach up the skill-ladder, resulting in the displacement of larger portions of the global workforce, including those in high-income countries. Nevertheless, it is imperative to acknowledge that the capacity of our planet to endure exploitation for the purpose of economic production is finite. Furthermore, it is crucial to recognize that the equitable distribution of any augmented productivity resulting from artificial intelligence cannot be guaranteed within society. To date, the phenomenon of automation has predominantly resulted in a redistribution of income and wealth from labor to capital owners, thereby exacerbating the growing disparity in wealth distribution worldwide. In addition, there remains a lack of understanding regarding the potential psychological and emotional ramifications of a society in which employment is scarce or obsolete. Moreover, there is a dearth of consideration given to the requisite policies and strategies that would be necessary to sever the link between unemployment and adverse health outcomes In accordance with the interpretation provided by McClelland [12].

In recent times, there has been a notable surge in global collaboration between robotics and artificial intelligence (AI) across diverse domains. The increasing popularity and prominence of robotics have contributed to the facilitation of daily living. Simultaneously, in the event that robots assume control over all occupations within the industrialized sector, there will be a decrease in human labor. Although robots enhance operational efficiency, they concurrently diminish employment opportunities. Robotic automation has assumed control over the entirety of blue-collar occupations. The integration of robots into professional occupations is currently being observed. Robots, which are artificially created entities, have the capability to engage in low-wage work during non-social periods, while simultaneously offering substantial solace to the global population. It is anticipated that forthcoming generations will potentially regard robots as instructors and caregivers once they acquire emotional capabilities such as compassion and advanced reaction detection. Thr gradual increase in human-robot interaction can be attributed to the fact that robots contribute to the enhancement of individual‘s lives, making them more convenient and comfortable [30].

2.2 Security of Machine Learning Algorithms

2.2.1 AI in Cyber Security

As cyber security threats are changing and developing constantly, immediate response is required and an automatic [88]. The significance of cyber security lies in its comprehensive scope, as it pertains to safeguarding our data from malicious cyber assailants who seek to illicitly acquire and exploit this information for detrimental purposes. Consequently, it is evident that they are susceptible to cyber attacks. A cyber attack refers to the deliberate initiation of an offensive action originating from one or multiple computer systems with the objective of either incapacitating the targeted computer, rendering its services inoperable, or illicitly obtaining access to the data stored within the targeted computer.

Artificial intelligence tools are frequently employed as a means to address the challenges posed by cyber threats. The utilization of artificial intelligence (AI) has proven to be advantageous for numerous organizations in enhancing their security measures and mitigating the likelihood of breaches. Machine learning and artificial intelligence play a crucial role in the realm of technology, particularly in the context of information security. These tools enable companies and individuals to effectively assess and analyze the various threats that may potentially jeopardize the integrity and security of an organization. With natural language processing, AI can provide greater predictive intelligence by skimming through articles, news, and research on cyber risks and curating material on its own [88]. The prevalence of cyber threats is increasing. Given the dynamic and evolving nature of cyber security threats, it is imperative to promptly and autonomously address them. Hence, the utilization of machine learning methods, particularly deep learning, which typically does not necessitate prior expertise or reliance on previous expert categorizations, holds significant significance in the deployment of artificial intelligence strategies for cyber security. Artificial intelligence (AI) systems are currently undergoing training to perform tasks such as malware identification, pattern recognition, and detection of subtle characteristics associated with malware or ransomware attacks. These tasks are accomplished through the utilization of sophisticated algorithms, which enable the AI systems to identify and prevent such threats from infiltrating computer systems. Through the utilization of natural language processing, artificial intelligence has the capability to enhance predictive intelligence by efficiently analyzing articles, news, and research pertaining to cyber risks. This process enables AI to autonomously curate relevant material.

The magnitude of these attacks would pose a significant challenge for the security personnel of a typical company. Consequently, a number of these threats will remain undetected and cause substantial harm to the network. In order to enhance operational efficiency and safeguard organizations against cyber threats, security professionals rely heavily on the assistance provided by advanced technologies, particularly artificial intelligence (AI) systems. Artificial intelligence (AI) possesses the capacity to effectively manage and process substantial volumes of data. The company's network experiences a significant amount of activity. A typical mid-sized firm experiences a significant volume of traffic. This suggests that a significant amount of data is regularly exchanged between customers and the company. It is imperative to ensure the protection of this information from individuals and software that may pose a threat. Nevertheless, cyber security experts encounter limitations in their ability to thoroughly examine all data for potential threats. Artificial intelligence (AI) represents a highly effective solution for identifying potential threats that are camouflaged within ordinary or routine behaviors. Due to its automated characteristics, this system possesses the capability to efficiently analyze substantial volumes of data and manage traffic. AI-driven technology, such as a personal proxy, has the potential to facilitate data transmission. Furthermore, it possesses the capability to detect and accurately identify potential hazards that may be concealed within a state of disorder. The presence of duplicative processes leads to a decrease in efficiency. As mentioned earlier, assailants often alter their tactics. However, the core security practices remain unchanged. If an individual is hired to perform these responsibilities, there is a possibility that they may experience boredom and consequently pose a risk to the security of your network. Artificial intelligence (AI) assumes responsibility for repetitive cyber security tasks, alleviating the burden on cyber security personnel, while emulating desirable human characteristics and excluding inherent limitations. It facilitates the identification and mitigation of inherent security vulnerabilities on a recurring basis.

Additionally, it conducts a comprehensive examination of the network infrastructure to identify potential security vulnerabilities that could pose risks to the network's integrity. The enhancement of detection and response times is observed. The initial stage in ensuring the security of a company's network involves the identification and detection of potential threats. It would be advantageous if one could promptly identify concerns such as the presence of unreliable data. Implementing this security measure will effectively safeguard your network from enduring detrimental consequences. The integration of artificial intelligence (AI) with cyber security represents a highly effective approach for the timely detection and response to cyber attacks.

For instance, the concept of authenticity protection is prevalent in the online landscape, as a considerable number of websites offer a user account functionality. This feature allows individuals to verify their identity, granting them access to a range of services and facilitating online transactions. Several websites use contact forms that require individuals to enter personal information. The implementation of supplementary security measures on websites containing confidential information and sensitive content is crucial for enterprises. The implementation of an upgraded security layer guarantees the protection of individuals accessing your network, so ensuring the safety of your guests. When a user seeks to establish a link with their account, artificial intelligence (AI) secures the authentication procedure. Within the domain of identification, artificial intelligence (AI) utilizes a wide array of methodologies, encompassing various approaches such as face recognition and fingerprint recognition, among others. The data related to these features can be employed to determine the authenticity of a login attempt.

Incidents of attacks during training sessions are more prevalent than commonly perceived. The majority of machine learning models in production undergo periodic retraining by incorporating new data into their systems. Social networks engage in ongoing analysis of user behavior, thereby affording each user the opportunity to influence the system by modifying their own behavior.

2.2.2 Attacks on Machine Learning Algorithms

Various types of attacks can be launched on machine learning (ML) models, contingent upon the attacker's specific objectives. attacks can occur at different stages of the ML pipeline, namely during the training and production phases. It is also possible to categorize these attacks as either targeting the underlying algorithm or the deployed model. At present, the most prevalent strategies employed are evasion, poisoning, and inference.

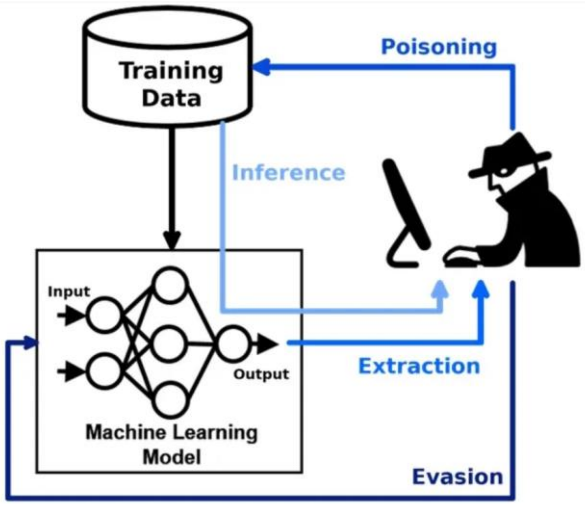


Figure 7 Attacks on ML [17]

Evasion Attack

The occurrence of assaults on artificial intelligence (AI) systems gives rise to a series of ethical ramifications that transcend the confines of technology. Fundamentally, these attacks are designed to intentionally manipulate AI models in order to generate inaccurate outcomes, which could have significant ramifications in practical scenarios. This phenomenon gives rise to apprehensions regarding the public's confidence in the dependability and impartiality of AI, thereby compromising its level of acceptance. Of particular concern is the unequal effect, as evasion attacks that succeed can disproportionately harm marginalized communities by capitalizing on biases inherent in AI systems. The deceptive characteristics exhibited by these attacks violate the ethical principle of transparency, thereby eroding the accountability of both AI developers and attackers. Furthermore, it is important to note that evasion attacks have the potential to violate individual privacy rights, necessitating a comprehensive comprehension of system vulnerabilities that could potentially expose confidential and sensitive data. Evasion attacks refer to deliberate attempts made by an attacker during the testing phase to modify input data in order to induce errors inside the machine learning system. Evasion attacks are characterized by their ability to exploit the blind spots and vulnerabilities of a system in order to induce certain errors, without directly modifying the system's behavior. The term perturbation pertains to the generation of an input that exhibits typical characteristics to a human observer, yet is erroneously categorized by machine learning models. One prevalent illustration involves the alteration of certain pixels within a photograph prior to its upload, thereby inducing a failure of the image recognition system to accurately classify the resultant image. Indeed, humans can be deceived by this adversarial entity. Various methods can be employed to execute this attack, depending on factors such as the model, dataset, and other relevant properties. The deliberate act of manipulating or deceiving the performance of an artificial intelligence (AI) system by introducing specific input data that is designed to exploit vulnerabilities or weaknesses in the system's learning algorithm. Evasion attacks refer to a category of adversarial attacks that are designed to induce AI models into generating inaccurate predictions or classifications. Evasion attacks can yield significant ramifications across diverse domains empowered by artificial intelligence, encompassing image recognition, natural language processing, and fraud detection [32].

Evasion attacks pertain to the manipulation of input data in a manner that introduces slight variations from the original data, thereby potentially inducing erroneous outcomes from an artificial intelligence model. As Papernot et al. highlight, evasion attacks can occur through techniques like adversarial examples, where imperceptible alterations to input data can lead to dramatically different model outputs [89]. This objective can be achieved through meticulous computation of noise or by altering the input. As an illustration an adversary possesses the capability to make subtle modifications to an image that has been accurately categorized by an artificial intelligence system for image recognition. The aforementioned alterations may elude human visual perception, yet they possess the capacity to induce significant misclassification of the image by the artificial intelligence system as reflected by the studies conducted by Polyakov [17].

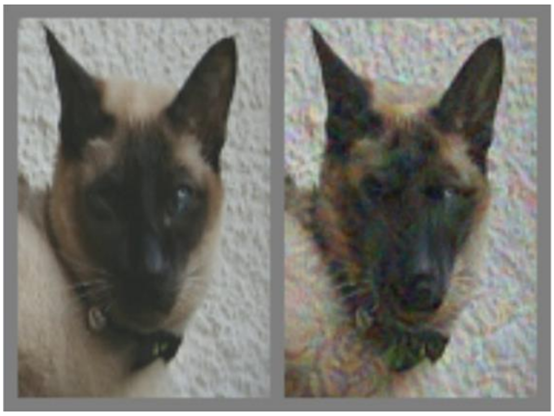


Figure 8 Evasion Attack [17]

Evasion attacks possess the potential to be employed with malicious intent, thereby facilitating fraudulent activities, dissemination of misinformation, or exerting adverse influence on the decision-making processes of AI systems. It is imperative to comprehend the ethical aspects associated with these attacks and to contemplate potential strategies for mitigation. They have the potential to deliberately induce incorrect outcomes in AI systems, thereby leading to significant real-world ramifications. In the context of medical diagnosis, the compromise of a diagnostic artificial intelligence (AI) system could have significant ramifications, such as the occurrence of misdiagnosis. This could subsequently result in the delay of appropriate treatment, posing potential risks to individual’s well-being and even their lives.

Privacy Attacks (Inference)

An inference attack refers to a sophisticated method that is employed to extract sensitive information from artificial intelligence models through the analysis of their outputs. In the context of this assault, adversaries leverage the discernible patterns and corresponding reactions exhibited by an artificial intelligence (AI) system to infer sensitive information pertaining to either the training data or the internal mechanisms of said system. Additionally, model inversion attacks, as described by Fredrikson et al. (2014), aim to reconstruct training data or uncover patterns within it by observing the model's outputs [90].

In terms of ethical considerations, inference attacks give rise to notable concerns as they undermine both user privacy and the security of confidential data. Through the process of reverse-engineering an artificial intelligence's outputs, malicious actors have the potential to extract information that was not explicitly included in the original data. This compromises the trust individuals have in the system's capacity to protect sensitive information. Furthermore, these attacks have the potential to compromise the confidentiality of proprietary or confidential data that is utilized for the purpose of training artificial intelligence models. The ethical considerations surrounding inference attacks are of utmost importance in guaranteeing the protection of data privacy, fostering user trust, and promoting the responsible implementation of AI technologies, particularly as AI systems become more involved in handling personal and sensitive information in various sectors including healthcare, finance, and law enforcement.

Membership Inference Attacks

Machine learning possesses the remarkable ability to transform diverse forms of data into mathematical representations. After the completion of training a machine learning model using various types of data such as images, audio, raw text, or tabular data, the outcome consists of a collection of numerical parameters. The attack methodology involves the training of an attack model with the objective of discerning the target model's response to the inputs used during training, as compared to its response to inputs that were not encountered during the training phase [91].

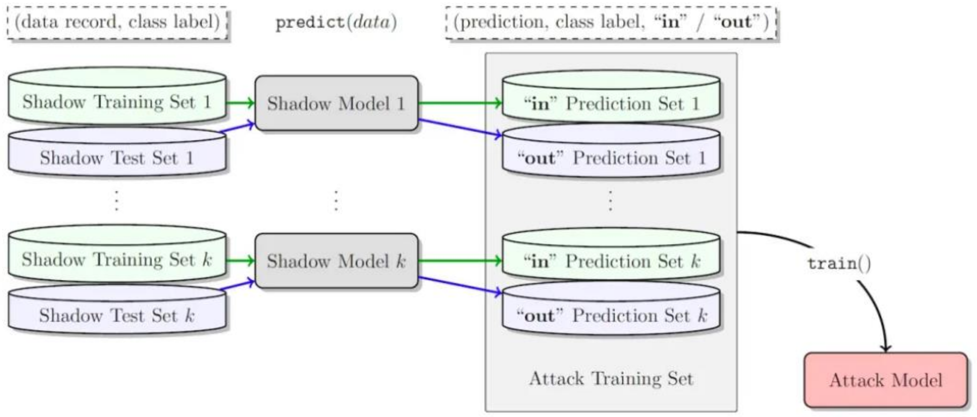
Membership inference attacks happen when an adversary has black-box access to a machine learning model and can observe its output on a given input. The adversary then uses this information to train an attack model to distinguish the target model's behavior on the training inputs from its behavior on the inputs that it did not encounter during training. The attack model can then be used to infer whether a particular data record was part of the model's training dataset or not [91]. In the majority of instances, the model becomes independent of the training dataset and utilizes the optimized parameters to classify new and unfamiliar instances into categories or make value predictions. In numerous instances, perpetrators can execute membership inference attacks without requiring access to the parameters of the machine learning model, solely by observing its output. The phenomenon of membership inference can give rise to significant security and privacy implications, particularly when the target model has been trained using sensitive data.

Every machine learning model possesses a collection of picked up parameters, the quantity and interconnections of which differ based on the specific algorithm and architecture employed. However, irrespective of the algorithm selected, all machine learning models undergo a comparable training process. The process commences by initializing the parameter values randomly, which are then iteratively adjusted to optimize the fit to the training data. Supervised machine learning algorithms, commonly employed in tasks such as image classification or spam detection, optimize their parameters to establish a mapping between input data and desired output predictions.

Machine learning models have a tendency to exhibit superior performance when evaluated on the data used for training. As an illustration, when incorporating additional images into the training dataset and subsequently subjecting them to the neural network, it becomes evident that the confidence scores assigned to the training examples surpass those assigned to previously unseen images. Membership inference attacks exploit this characteristic to uncover or reconstruct the instances employed for training the machine learning model. The utilization of individual’s data records for training the model may have potential privacy implications. In the context of membership inference attacks, it is not a strict requirement for the adversary to possess explicit knowledge regarding the internal parameters of the targeted machine learning model. In contrast, the perpetrator possesses knowledge solely pertaining to the algorithm and architecture of the model, such as Support Vector Machines (SVM) or neural networks, or the specific service employed in the model's creation. The proliferation of machine learning as a service (MaaS) solutions provided by prominent technology companies like Google and Amazon has led to a significant inclination among developers to utilize these offerings rather than constructing their own models from the ground up.

The developer is solely responsible for configuring a novel model and furnishing it with a dataset for training purposes. The remaining tasks are handled by the service. One potential tradeoff arises when the attackers possess knowledge regarding the specific service utilized by the victim, as this enables them to employ the same service in order to construct a membership inference attack model. During the 2017 IEEE Symposium on

Security and Privacy, scholars from Cornell University presented a membership inference attack technique that demonstrated efficacy across prominent cloud-based machine learning platforms [19].



Membership inference attacks observe the behavior of a target machine learning model and predict examples that were used to train it.

Figure 9 Membership Attacks [19]

This technique involves the creation of arbitrary records by an adversary targeting a machine learning model hosted on a cloud-based service. The perpetrator inputs each individual record into the model. The attacker adjusts the features of the record and reevaluates it using the model, taking into account the confidence score provided by the model. The iterative procedure persists until the model attains a significantly elevated level of confidence. Currently, the record bears a resemblance, if not an exact match, to one of the instances employed for training the model.

Once a sufficient number of records with high confidence have been collected, the perpetrator utilizes the dataset to train a collection of "shadow models" with the objective of predicting whether a given data record was included in the training data of the target model. This process generates a collection of models that are capable of training a membership inference attack model. The ultimate model has the capability to make predictions regarding the inclusion of a data record in the training dataset of the target machine learning model.

The study conducted by the researchers revealed that the aforementioned attack demonstrated a high degree of success across various machine learning services and architectures. The results of their study indicate that a proficiently trained attack model possesses the ability to discern between individuals belonging to the training dataset and those who do not, while also assigning a high level of confidence to the predictions made by the target machine learning model.

Poisoning

Within the realm of Artificial Intelligence (AI), a poisoning attack refers to the deliberate act of introducing malevolent or manipulated data into the training set of an AI model. The primary objective of such an attack is to compromise the learning process of the model and subsequently impair its overall performance. The objective of this attack is to subtly manipulate the underlying data used for AI training, leading to inaccurate predictions or classifications.

Poisoning attacks give rise to significant ethical concerns due to their potential to undermine the integrity of AI systems and the reliability of their outputs. Attackers exploit vulnerabilities in the learning process of AI models by deliberately contaminating the training data, thereby undermining the trust that users have in the objectivity and reliability of AI. The manipulation of AI's knowledge base can potentially have significant implications across various domains, ranging from the potential for inaccurate medical diagnoses to the potential compromise of safety in autonomous vehicles. Furthermore, the utilization of poisoning attacks has the potential to introduce subtle biases into artificial intelligence systems, thereby amplifying concerns pertaining to the principles of fairness and equity. Given the growing significance of AI in crucial decision-making procedures, it is imperative to acknowledge and examine the ethical aspects associated with poisoning attacks. This examination is necessary to guarantee that AI technologies fulfil their potential of improving human welfare and advancing society.

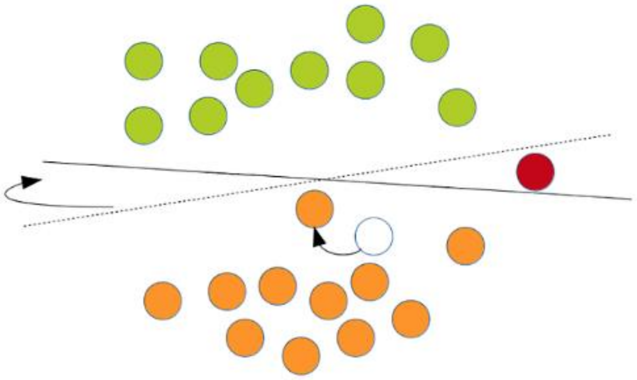


Figure 10 Poisoning Attacks [17]

Split-view poisoning is a sophisticated strategy that capitalizes on the discrepancy between the data used for curation and the data used for training the artificial intelligence (AI) model. The magnitude of this discrepancy can be significant and unpredictable, thereby enhancing the efficacy of the technique [17]. Similarly, in the event that threat actors are able to obtain access to training data, they possess the capability to manipulate the artificial intelligence (AI) and machine learning (ML) configuration, causing it to classify benign software data as malicious. Moreover, threat actors have the potential to exploit the training data by creating a backdoor. *"Hackers may use AI to help choose which is the most likely vulnerability worth exploiting. Thus, malware can be placed in enterprises where the malware itself decides upon the time of attack and which the best attack vector could be. These attacks, which are, by design, variable, make it harder and longer to detect."* says Grajek [37].

An alternative method of engaging in data poisoning involves enhancing the intelligence of malware. Malicious actors employ such malware in order to compromise email systems by replicating phrases in order to deceive the underlying algorithm. The phenomenon under consideration extends to the field of biometrics, wherein malicious actors impede the access of authorized users while surreptitiously gaining entry. Moreover, the phenomenon commonly referred to as "fake news" can be classified as a form of data poisoning. The efficacy of popular social media algorithms in detecting and addressing fake news is either compromised due to technical limitations or influenced by corrupt practices. Frequently, there is an observable trend wherein inaccurate information supplants genuine news on a user's news feed.

2.3 Solutions for Building a Secure AI

Artificial Intelligence (AI) has emerged as a disruptive phenomenon, holding the potential to fundamentally reshape various sectors such as industries, healthcare, finance, and beyond. With the growing integration of AI systems into our daily routines, it is imperative to prioritize the maintenance of their security and safety. This essay highlights the paramount importance of constructing AI systems that are secure and safe, with a particular focus on their pivotal role within the ethical, societal, and technological domains. The primary significance of ensuring secure artificial intelligence (AI) resides in the protection of sensitive data and the preservation of privacy. Artificial intelligence (AI) systems frequently handle substantial quantities of personal and confidential data, rendering them attractive targets for cyberattacks. In the absence of comprehensive security measures, these systems have the potential to serve as vulnerable points of entry for malevolent individuals aiming to gain unauthorized access to confidential information. The assurance of artificial intelligence (AI) security is not solely a matter of legal and ethical responsibility, but also a fundamental entitlement of individuals.

Furthermore, the implementation of safe artificial intelligence (AI) is crucial in safeguarding the well-being and preservation of human lives. In industries such as autonomous vehicles and healthcare, the implementation of AI-driven decision-making processes carries significant implications in practical contexts. The assurance of AI system safety, particularly in applications with critical safety implications, is imperative to mitigate the occurrence of accidents, injuries, and fatalities. The implementation of fail-safe mechanisms in AI algorithms and systems is considered both a moral obligation and a legal mandate in numerous jurisdictions. Moreover, the implementation of robust and reliable artificial intelligence (AI) systems plays a crucial role in cultivating trust and garnering user acceptance. The likelihood of users embracing and adopting AI systems is positively influenced by their perception of reliability and security in these technologies. The establishment of trust in artificial intelligence (AI) is of paramount importance in order to effectively incorporate it into diverse sectors and to maximize its positive impact on society as a whole.

In addition, the establishment of robust artificial intelligence (AI) systems is imperative for safeguarding national security and enhancing economic competitiveness. As artificial intelligence (AI) increasingly becomes an essential component of crucial infrastructure, such as defence and financial systems, it is important to acknowledge that there exist potential vulnerabilities that can be exploited by adversarial entities. The assurance of security in artificial intelligence (AI) systems is therefore a matter of significant strategic importance for nations. The significance of constructing artificial intelligence systems that are both secure and safe is of utmost importance. The protection of privacy, prevention of harm, cultivation of trust, assurance of national security, and adherence to ethical standards are all important considerations. In light of the ongoing advancements and transformative impact of artificial intelligence (AI) on our global landscape, it becomes increasingly imperative to accord utmost importance to the domains of security and safety. This imperative extends beyond mere technical requirements, encompassing moral, legal, and societal considerations. The prioritization of secure and safe AI systems development, with a focus on their positive contribution to the dynamic technological landscape, falls under the shared responsibility of AI developers, researchers, and policymakers.

2.3.1 Building a Trust Worthy AI

The importance of trustworthiness in artificial intelligence (AI) cannot be overstated due to a number of significant factors. Firstly, it facilitates the increase in user acceptance and adoption. Individuals are more inclined to adopt artificial intelligence (AI) solutions when they hold the belief that these systems function in an ethical, fair, and consistent manner. The establishment of trust in artificial intelligence (AI) instils a sense of assurance in its dependability, thereby cultivating a favourable user encounter. Furthermore, trust plays a crucial role in fostering transparency and accountability. Trustworthy artificial intelligence (AI) systems possess the capability to offer justifications for their decisions and actions, thereby guaranteeing transparency in their operational processes. The transparency of AI systems enables users and stakeholders to effectively hold them accountable, especially in sectors where decisions carry significant implications.

Moreover, trust plays a crucial role in mitigating AI bias. The presence of bias in artificial intelligence (AI) systems, which can stem from biased training data or algorithms, has the potential to result in discriminatory consequences. The establishment of trustworthiness in artificial intelligence (AI) systems necessitates the implementation of strong mechanisms that can effectively detect and address bias, thereby guaranteeing fairness and impartiality. The concept of trust exhibits a strong correlation with the ethical advancement of artificial intelligence. The prioritization of user well-being, the respect for privacy, and the adherence to ethical guidelines should be the primary considerations in the development and implementation of ethical artificial intelligence (AI) systems. Trustworthy artificial intelligence (AI) acts as a motivating factor for developers to maintain ethical standards during the entire process of AI development. The development of reliable artificial intelligence (AI) systems is not solely driven by technical requirements, but also by ethical and societal obligations. The successful integration of AI into society is contingent upon the establishment of trust, as it contributes to user acceptance, promotes transparency and accountability, guarantees fairness, and upholds ethical standards. In the context of AI's ongoing transformation of our society, the imperative of prioritizing trustworthiness emerges as a crucial factor in guaranteeing that AI systems effectively cater to the collective welfare of humanity.

A document was authored by the High-Level Expert Group on Artificial Intelligence (AI HLEG). The individuals comprising the AI High-Level Expert Group (AI HLEG) expressed their endorsement of the comprehensive structure for Trustworthy AI outlined in these Guidelines, while acknowledging that they may not fully concur with each individual assertion made within the document [10]. The paper emphasises that a trustworthy artificial intelligence (AI) encompasses three fundamental elements that must be upheld throughout the entirety of the system's life cycle. Firstly, it is imperative that the AI system operates within the boundaries of the law, adhering to all relevant laws and regulations. Secondly, the AI system must embody ethical principles and values, ensuring its conduct aligns with established ethical standards. Additionally, it is imperative for an AI system to possess robustness, encompassing both technical and social aspects. This is crucial due to the potential of AI systems to inadvertently inflict harm, despite being developed with good intentions. Each individual component possesses a level of necessity, yet lacks sufficiency in isolation, for the attainment of Trustworthy AI. Ideally, the three components function synergistically and exhibit overlapping functionality. If, in practical application, conflicts arise between these components, society should strive to harmonize them.

The proposed Ethics Guidelines for Trustworthy AI provide a comprehensive set of measures, encompassing both technological and non-technical approaches, to effectively enforce the prerequisites for the realization of Trustworthy AI. The technical approaches encompass the integration of Trustworthy AI requirements throughout the various stages of an AI system's lifecycle, including design, development, and utilization. These requirements are assessed and implemented continuously, while also leveraging established solutions such as equity-by-design in supervised machine learning techniques, algorithmic repeatability, resilience against bias and corruption, and the creation of causal models. Non-technical approaches encompass the establishment of a team characterized by diversity and multidisciplinary expertise, the dissemination of explicit and proactive communication to stakeholders regarding the capabilities and limitations of the artificial intelligence (AI) system, and the guarantee of traceability throughout the AI system's operations. The present guidelines establish a comprehensive framework for the attainment of AI systems that are characterized by trustworthiness. The framework does not explicitly address the initial component of Trustworthy AI, which pertains to the legality of AI systems [10].

The objective of this initiative is to provide guidance pertaining to the fostering and securing of ethical and robust artificial intelligence (AI) systems. These Guidelines are intended for all stakeholders and aim to surpass a mere enumeration of ethical principles. Instead, they offer direction on how these principles can be put into practice within socio-technical systems. The paper concludes by presenting examples of opportunities and critical concerns that arise from the use of AI systems [10]. Europe’s early start towards AI regulation offers an opportunity to establish an effective legal framework, grounded in the rule of law. But beyond legislative frameworks at the level of nation states, multilateral public-private partnership is needed to ensure AI governance can have an impact today, not just a few years from now, and at the international level.[66]. A detailed description on how to achieve a governed AI that has been produced by Microsoft is explained in appendix A.

solutions against discrimination

The ability to effectively confront the ethical dilemmas presented by artificial intelligence (AI) is rapidly becoming an essential requirement for effective governance as interpreted by HLEG . Regrettably, the existing mechanisms designed to supervise human decision-making frequently prove inadequate when extended to the realm of artificial intelligence (AI). Therefore, it is imperative to develop novel mechanisms that can effectively guarantee the ethical alignment of artificial intelligence (AI) systems, which are becoming increasingly prevalent in various aspects of society.

According to Floridi and Cowls, prominent organizations within the realms of politics, business, and academia have recognized the pressing need for addressing this issue and have consequently developed ethical guidelines pertaining to reliable artificial intelligence. Nevertheless, the implementation of these guidelines remains discretionary. Furthermore, it is evident that the industry is deficient in effective tools and incentives that can facilitate the translation of ethics principles at a high level into criteria that are both verifiable and actionable in the context of designing and implementing artificial intelligence.

Several recent developments suggest that ethics-based auditing is a promising approach to addressing the disparity between ethical principles and practical implementation in the field of AI ethics. In a seminal publication released in April 2020, esteemed scholars affiliated with prominent institutions such as Google, Intel, Oxford, Cambridge, and Stanford propose the involvement of independent auditors in evaluating the veracity of assertions pertaining to safety, security, privacy, and fairness put forth by developers of artificial intelligence (AI) systems. Concurrently, prominent professional services firms such as PwC and Deloitte are actively constructing frameworks aimed at assisting clients in the creation and implementation of reliable artificial intelligence [22]. This development is highly encouraged. However, it is crucial to maintain a realistic perspective regarding the achievable outcomes of ethics-based auditing of AI. Ethics-based auditing serves as a governance mechanism employed by organizations involved in the design and implementation of AI systems, with the aim of regulating and exerting influence over the conduct of said systems. Ethics-based auditing is distinguished by a methodical procedure through which the conduct of an organization is evaluated to determine its alignment with pertinent principles or norms [22]. The implementation of ethics-based auditing plays a significant role in promoting and upholding principles of good governance. Similar to how businesses rely on physical infrastructures for their success, the flourishing of interactions between agents necessitates the presence of an ethical infrastructure.

To attain Trustworthy AI, it is imperative to foster inclusivity and diversity across all stages of the AI system's life cycle. In addition to the inclusion and engagement of all relevant stakeholders throughout the entirety of the procedure, it is imperative to guarantee equitable access through inclusive design methodologies and equitable treatment. This requirement is intricately connected with the principle of fairness.

Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition, such as the homogenisation of prices by means of collusion or a nontransparent market. Identifiable and discriminatory bias should be removed in the collection phase where possible. The way in which AI systems are developed (e.g. algorithm‘s programming) may also suffer from unfair bias. This could be counteracted by putting in place oversight processes to analyse and address the system’s purpose, constraints, requirements and decisions in a clear and transparent manner. Moreover, hiring from diverse backgrounds, cultures and disciplines can ensure diversity of opinions and should be encouraged [10]. The collection of specific data categories, such as ethnicity data, is frequently advantageous or even essential in order to conduct audits on artificial intelligence systems for the purpose of identifying and addressing instances of discrimination [21]. Moreover, Utilizing a bias impact assessment can be extremely beneficial in identifying the primary stakeholders, their respective interests, and their relative positions of influence when obstructing or allowing essential changes, as well as the immediate and long-term consequences of these actions. Dwork et al. [84] initially proposed the concept of fairness through awareness. According to this theory, in order to address or eliminate bias in an algorithmic model, it is necessary to first recognize and comprehend the biases and their root causes. The bias impact assessment is significant because it serves the purpose of evaluating bias.

In addition to facilitating a deeper understanding of complex social and historical contexts, the assessment can also strengthen a pragmatic and ethical framework for identifying potential biases. Consequently, it can aid in the identification and reduction of these obstacles during the evaluation of AI systems that make automated decisions, thereby promoting accountability. Various ethical categories can be evaluated using qualitative and quantitative measures. These categories encompass a variety of ethical concerns, including prejudice, privacy, security, and mental health.

2.3.2 Building a Secure AI

Artificial intelligence (AI) possesses significant potential for constructing a more advanced and intelligent society. However, it concurrently encounters substantial security vulnerabilities. As a result of insufficient security considerations during the initial stages of AI algorithm development, malicious actors possess the capability to manipulate the outcomes of inference, thereby causing erroneous assessments. Security risks in critical domains such as healthcare, transportation, and surveillance can have severe and detrimental consequences. Instances of AI systems being compromised can lead to significant financial losses and pose a threat to individual's physical well-being. Security will always need the power of humans and machines, and more powerful AI automation will help us optimize where we use human ingenuity. The more we can tap AI to render actionable, interoperable views of cyber risks and threats, the more space we create for less experienced defenders who may just be starting their careers. In this way, AI opens the door for entry-level talent while also freeing highly skilled defenders to focus on bigger challenges [64]. The presence of AI security risks is evident not only in theoretical analyses but also in the practical implementation of AI systems. For example, individuals with malicious intent have the ability to create files that can evade detection by artificial intelligence-based tools, or introduce extraneous elements into voice commands used for smart home voice control, thereby triggering the execution of harmful applications. In addition to manipulating data returned by a terminal, attackers have the ability to intentionally engage in malicious conversations with a chat robot, thereby inducing a prediction error in the underlying artificial intelligence system. I would like to request that the user's text be rewritten in an academic manner.

AI is making significant strides in various deployment scenarios, including but not limited to robots, virtual assistants, autonomous driving, intelligent transportation, smart manufacturing, and Smart Cities, leading to noteworthy achievements. Prominent corporations, including Google, Microsoft, and Amazon, have embraced artificial intelligence (AI) as a central component of their long-term growth strategies. In the year 2017, Google's DeepMind introduced AlphaGo Zero, a program that improved its ability to play the game of Go solely through self-play. Remarkably, AlphaGo Zero managed to surpass the previous champion-defeating version of AlphaGo in just three days of self-training. AlphaGo Zero demonstrates the capacity to acquire novel insights and formulate unconventional strategies, thereby highlighting the significant possibilities associated with the application of artificial intelligence (AI) in transforming human existence [36]. Prevention and detection are considered the most effective strategies for mitigating data poisoning. The monitoring of production models enables the detection of data or concept drift, which encompasses phase shifts, imbalances, and alterations in data density and distribution. By implementing a system of continuous monitoring, we can effectively mitigate the risk of the model being influenced by compromised data according to Excella [35].

To mitigate these AI security risks according to Huawei [36], AI system design must overcome five security challenges:

- Software and hardware security: The code of applications, models, platforms, and chips may have

vulnerabilities or backdoors that attackers can exploit. Further, attackers may implant backdoors in models to

launch advanced attacks. Due to the inexplicability of AI models, the backdoors are difficult to discover.

- Data integrity: Attackers can inject malicious data in the training stage to affect the inference capability of AI

models or add a small perturbation to input samples in the inference stage to change the inference result.

- Model confidentiality: Service providers generally want to provide only query services without exposing the

training models. However, an attacker may create a clone model through a number of queries.

- Model robustness: Training samples typically do not cover all possible corner cases, resulting in the

insufficiency of robustness. Therefore, the model may fail to provide correct inference on adversarial examples.

- Data privacy: For scenarios in which users provide training data, attackers can repeatedly query a trained model to obtain user’s private information.

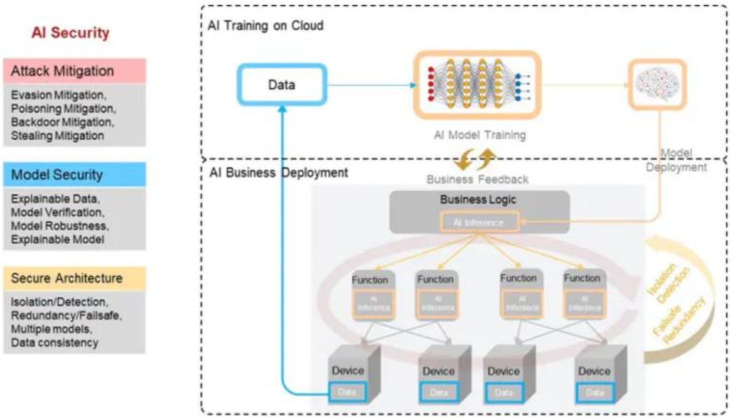


Figure 11 AI Security Defence Architecture [61]

For the Mitigation of attacks, it is advised to Develop defensive mechanisms to counter known attack strategies. Common AI security attacks encompass evasion attacks, poisoning attacks, backdoor attacks, and model extraction. Numerous countermeasures have been proposed in academic literature to address these attacks. These countermeasures include adversarial training, Network Distillation, adversarial detection, DNN model verification, data filtering, ensemble analysis, model pruning, PATE, and others [36]. In specific scenarios, adversarial training has proven to be an effective approach for defending models. This defence technique involves augmenting the training data of a supervised model with adversarial examples, thereby enhancing the model's ability to accurately classify such instances. The objective of this approach is to mitigate the potential harm caused by adversarial examples through the utilization of a training process that incorporates both clean and adversarial data Based on the available evidence from Sandia National Laboratories [39]. There are various methodologies that can be employed to implement adversarial training. The most direct method entails integrating adversarial examples alongside their corresponding "accurate" labels, thereby strengthening the model against specific attacks and striving to maintain model precision [39].

The prevalence of adversarial machine learning poses significant challenges to model security. Evasion attacks, poisoning attacks, and various techniques that exploit vulnerabilities and backdoors are not only effective but also possess significant transferability, thereby posing substantial risks of misjudgement by AI models. Therefore, in addition to safeguarding against established forms of attack, it is imperative to enhance the security of an artificial intelligence (AI) model itself in order to mitigate the potential harm inflicted by alternative attack vectors. Possible techniques encompass model detectability, model verifiability, and model explain-ability [36]. The field of AI model security encompasses a diverse array of techniques that are specifically developed to ensure the protection of data and models, with a focus on maintaining their confidentiality, integrity, and availability.

According to OWASP, the mitigation of these risks can be achieved through several measures, such as safeguarding the AI pipeline from data poisoning or AI supply chain assaults, concealing model parameters when feasible, implementing throttling and monitoring mechanisms for model access, identifying and addressing specific input manipulation, and including considerations of these attacks throughout the model training process [67]. A detailed guideline could be found in Appendix 6.1.2. The process of ensuring the security of AI models commences by implementing measures to protect the data that these models engage with. This process entails the implementation of advanced algorithms such as AES-256 for encrypting sensitive data, ensuring secure storage of the data, and utilizing secure channels for transmitting the data. Furthermore, it is imperative to implement rigorous access controls in order to restrict the individuals who are authorized to access confidential information. For example, in the context of a company utilizing an artificial intelligence (AI) model for the purpose of customer service, it becomes necessary for the model to have requisite access to sensitive customer data, encompassing details like contact information and purchase history. The implementation of strong encryption measures will effectively safeguard the data, rendering it indecipherable in the event of interception, unless the appropriate encryption key is possessed [40].

An informed user base is considered to be a highly effective means of defence against security threats. It is of utmost importance to provide users with comprehensive education regarding the potential risks that may arise from the utilization of AI models, including but not limited to phishing attacks and the improper manipulation of deepfakes. Regular training sessions have the potential to assist users in recognizing potential security threats and comprehending the appropriate protocols for reporting such threats to the relevant authorities. Within the realm of AI model security, the principles governing the design of secure and private AI encompass various aspects. These include the practice of data minimization, the incorporation of privacy as a fundamental aspect of design both by default and by design, the implementation of secure data handling protocols, the acquisition of user consent and the provision of transparency, the utilization of anonymization and pseudonymization techniques, and the integration of privacy-preserving technologies.

The establishment of software security architecture is an essential component in the process of designing and constructing artificial intelligence (AI) systems. This practice serves to safeguard the integrity of data, models, and algorithms by preventing unauthorized access, manipulation, or misuse. The field of architecture security necessitates a meticulous examination of the potential security hazards associated with the development of artificial intelligence (AI) systems. It is imperative to enhance preventive measures and impose restrictive conditions to mitigate risks and guarantee the secure, dependable, and manageable advancement of AI. When utilizing AI models, it is imperative to conduct a thorough analysis and assessment of the potential risks associated with their implementation. This assessment should be based on the unique characteristics and architecture of the specific services being utilized. Subsequently, it is essential to devise a comprehensive AI security architecture and deployment solution that incorporates various security mechanisms such as isolation, detection, failsafe, and redundancy. [36]

The initial phase in acquiring and enhancing proficiency in software security architecture for AI systems involves the identification of the security requisites pertinent to the project at hand. This entails conducting an analysis of the potential threats, risks, and vulnerabilities that your artificial intelligence (AI) system may encounter, in addition to the regulations, standards, and best practices that must be adhered to [44].

The subsequent phase entails the selection of suitable security patterns that align with the security prerequisites of your AI system and its architectural framework. Security patterns are a set of reusable solutions that aim to tackle prevalent security issues encountered in the process of software design and implementation. Given the interdisciplinary nature of developing machine learning (ML) applications, it is imperative to address any knowledge disparities regarding security among team members during the initial design phase.

The fundamental concept underlying the security design pattern is the segregation of input from features. In order to train a model, it is necessary to extract features from the raw input. However, it is important to note that in the majority of machine learning problems, inputs are not utilized directly as features [46]. Various transformations, such as standardization, scaling, and encoding, are commonly employed on the input data in order to convert them into suitable features that can be utilized as inputs for algorithms. However, if these transformations are performed as preprocessing steps, reproducing them during prediction becomes problematic as the input for prediction will remain in its original, unprocessed format. It is crucial to establish a clear distinction between the inputs and features, encapsulate the preprocessing procedures, and incorporate them into the model to guarantee reproducibility [46].

When it comes to privacy, from a practical standpoint, it is advisable to limit the availability of sensitive data and generate anonymized duplicates for purposes that are not compatible, such as analytics. It is imperative to establish and record a purpose or lawful basis prior to data collection, and effectively convey this purpose to the user in a suitable manner. The practice of storing aggregated personal data in highly secure and limited-access locations, as well as the decentralization of machine learning processes to eliminate the necessity of consolidating data in a singular location. In contrast, the model undergoes training through several iterations conducted at various locations could lead to privacy protection of data [67].

2.3.3 Ethical, Legal and Social Future of AI

Artificial intelligence (AI) has garnered considerable attention and has been the subject of extensive scholarly investigation over the course of numerous decades, culminating in notable progress in recent times. Artificial intelligence (AI) possesses the capacity to profoundly transform society. However, apprehensions regarding its societal implications, encompassing the economy, healthcare, education, and employment, are increasingly mounting. The field of Artificial Intelligence (AI) has had a profound impact on the field of human psychology, leading to significant changes in our cognitive processes, behaviours, and emotional reactions. The pervasive presence of artificial intelligence (AI)-driven technologies has fundamentally transformed our understanding and engagement with the world.

A notable transformation can be observed within the domain of human cognition. Artificial intelligence (AI)-enabled technologies have become deeply embedded in our everyday routines, exerting a significant impact on our cognitive processes. The utilization of AI-powered search engines and recommendation systems has brought about alterations in our attention spans, critical thinking capacities, and information processing methods. The stakes are high. AI may well represent the most consequential technology advance of our lifetime. And while that's saying a lot, there's good reason to say it. Today's cutting-edge AI is a powerful tool for advancing critical thinking and stimulating creative expression. It makes it possible not only to search for information but to seek answers to questions. It can help people uncover insights amid complex data and processes. It speeds up our ability to express what we learn more quickly. Perhaps most important, it's going to do all these things better and better in the coming months and years [65]. Moreover, artificial intelligence (AI) has significantly influenced and shaped human behaviour and social interactions. Social media platforms utilize artificial intelligence algorithms to curate personalized content, which can potentially contribute to the reinforcement of pre-existing beliefs and the development of polarized viewpoints. The aforementioned modifications carry substantial ramifications for fostering open dialogue and cultivating empathetic comprehension among individuals.

AI-driven virtual assistants and chatbots have significantly transformed our social interactions, eliciting individuals to form emotional connections with these virtual entities, despite their inherent absence of authentic emotions. The aforementioned phenomenon prompts significant inquiries regarding the convergence of technology and human emotions. In brief, the influence of artificial intelligence (AI) on human psychology is significant and diverse, encompassing various aspects such as cognition, behaviour, and emotional experiences. The comprehension and management of psychological impacts assume a crucial role in establishing a harmonious and mutually advantageous association between humans and technology, as the field of AI progresses.

The advancement and progression of this technology are occurring at an accelerated rate. Nevertheless, the process was not as seamless and effortless as it appeared to us. The current stage of artificial intelligence has been achieved through the collective efforts of numerous individuals over the course of several years, involving extensive dedication and diligent work. Artificial intelligence (AI), as an innovative technology, is subject to numerous debates and controversies regarding its future implications and potential effects on humanity. While there are potential risks involved, this situation also presents a significant opportunity. Artificial intelligence (AI) is expected to be utilized for the purpose of augmenting defensive and offensive cyber operations. Furthermore, novel methods of cyber-attack are anticipated to be developed in order to exploit specific vulnerabilities inherent in AI technology.

The proliferation of narrow artificial intelligence (AI) technology has led to the evaluation of most AI systems based on a "task-oriented evaluation" approach as mentioned by Hernández-Orallo [94]. This evaluation method assesses the AI's performance relative to specific tasks and measurable outcomes [48]. The advancements observed in these evaluations indicate that AI is increasingly valuable; however, they do not necessarily imply an increase in AI's level of intelligence. The process of measuring and evaluating artificial intelligence necessitates the application of classification and comprehension pertaining to the principal technologies that are influencing the domain. The field of artificial intelligence (AI) exhibits a wide range of diversity and is experiencing rapid growth, making it resistant to straightforward categorization. Russell and Norvig (2016) conducted a study that was published in the Journal of Artificial Intelligence Research (JAIR)[49]. Their research indicates that artificial intelligence (AI) is undergoing a transformation from rule-based systems to machine learning. This shift enables machines to acquire knowledge and enhance their performance through experiential learning. This implies that artificial intelligence (AI) is progressively becoming more versatile and experiencing an expansion in its functionalities. Moreover, the authors posit that artificial intelligence (AI) is poised to revolutionize numerous sectors.

The establishment of secure environments for educational and exploratory purposes, focused on the potential enhancements of artificial intelligence in public service, is imperative. Furthermore, these initiatives should facilitate the seamless transition of these activities into large-scale implementations across the public service sector. The absence of safe spaces and the potential absence of high-priority objectives may lead to the politicization of any perceived failures or lack of initial success, thereby impeding the ability of representative governments to effectively respond to the swiftly advancing era of artificial intelligence. A potential concern regarding the future of artificial intelligence (AI) in the realm of public service involves the utilization of AI by governmental entities, whether with good intentions or not, for the purpose of surveillance and monitoring of individual's activities. Rather than facilitating the empowerment of individuals, artificial intelligence (AI) is employed to categorize and screen behaviours that are not sanctioned by the government [50]. The dissemination of information regarding the activities of AI and its algorithms for governmental purposes is not transparent to the general public. Furthermore, the public remains unaware of the existence of distinct risk, credit, and behavioural scores assigned to individuals, which subsequently impact their societal privileges and limitations.

The utilization of natural language processing and machine learning techniques in platforms like ChatGPT is significantly transforming the domain of consumer content. These advancements facilitate the creation of dynamic and interactive experiences, which were previously only achievable through human curation. The transition from the curation of consumer-generated content to the creation facilitated by models such as ChatGPT has resulted in a profound and impactful alteration in the processes of content generation and consumption. The adoption of AI-driven content creation presents several benefits, such as the capacity to produce customized responses, manage large quantities of inquiries, and adjust to specific user preferences [51].

3 Requirements / Methods

The proliferation and incorporation of artificial intelligence technologies into diverse facets of human existence in the era of AI give rise to intricate ethical, social, and legal quandaries. The present study aims to investigate the primary research issue at hand.

In light of the rapid proliferation of artificial intelligence (AI) technologies in various fields, it is important to examine the complex ethical issues, societal ramifications, and legal factors that arise. Furthermore, it is crucial to establish a comprehensive framework that promotes responsible development and utilization of AI.

This research problem highlights the necessity of conducting a comprehensive investigation and analysis of the complex interaction between the potential advantages of AI and the corresponding challenges it poses. The study will explore the diverse aspects of ethical decision-making, societal change, and the developing legal framework in relation to the implementation of artificial intelligence. Furthermore, the study will investigate practical suggestions and principles that can assist policymakers, practitioners, and stakeholders in effectively navigating the intricate realm of AI ethics, social consequences, and legal adherence.

The objective of this chapter is to provide an overview of the systematic approach and research methods utilized in the examination, analysis, and resolution of the complex challenges presented by the widespread adoption of artificial intelligence (AI) technologies. This chapter functions as an essential resource for comprehending the methodology employed in the research and the manner in which the research inquiries pertaining to ethical, social, and legal concerns in the era of artificial intelligence are being tackled.

The inclusion of a methodology chapter in a research study serves to enhance transparency and rigor, enabling readers to assess the credibility and validity of the study's findings. Establishing the credibility of the research and demonstrating the appropriateness of the chosen research methods are crucial for addressing the intricate ethical, social, and legal concerns associated with artificial intelligence (AI).

3.1 The Employment of the Research Methodologies

This section explores the research methodologies and techniques that have been carefully selected to conduct a thorough investigation into the ethical, social, and legal complexities inherent in the field of artificial intelligence (AI) technologies. The integration of various qualitative research methods, such as literature review, content analysis, and interviews, has been utilized to facilitate a comprehensive examination of the research problem.

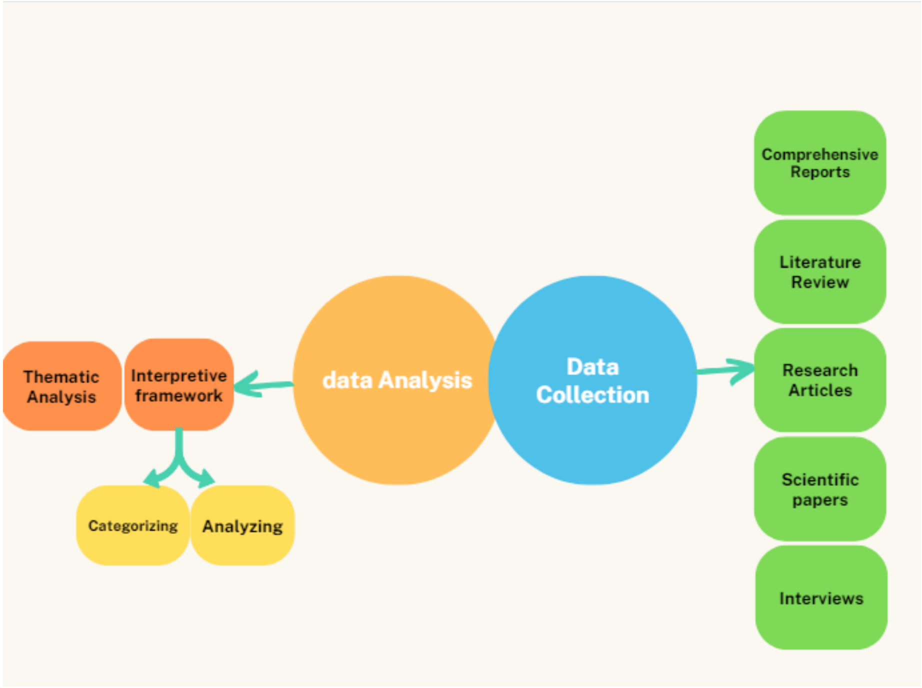


Figure 12 Research Approaches

3.1.1 Employed Methods for Data Collection

The qualitative research approach has been adopted in light of the complex and diverse nature of the subject matter. This methodology has been acknowledged as suitable for exploring complex phenomena and capturing the contextual essence. The primary aim of this study has been to gain in-depth understanding of the ethical, social, and legal dimensions of artificial intelligence (AI). This has been achieved through a comprehensive analysis of existing scholarly literature and the collection of firsthand data from interviews.

The initiation of data collection originated from an extensive literature review, which involved a thorough examination of scientific papers, research articles, comprehensive reports, and relevant policy documents. The collection of scholarly sources played a crucial role in addressing the ethical, social, and legal challenges associated with artificial intelligence (AI). Additionally, these sources provided valuable insights into the prominent themes and future directions within the vast field of AI. In addition, the utilization of secondary data obtained from reliable academic databases and platforms relevant to the research field enhanced the ability to identify significant patterns, trends, and insights pertaining to the complex dynamics of artificial intelligence.

The research process encompassed two distinct approaches to gathering data, namely descriptive and experimental methods. The former encapsulated an immersive and extensive scrutiny of scholarly literature. This process enabled the identification and analysis of significant ethical dilemmas, societal consequences, and legal complexities that arise from the deployment of artificial intelligence. The latter, which refers to the experimental arm, was utilized to analyze and examine specific scenarios or use cases in which AI is integrated, with the aim of comprehensively assessing its operational dynamics within carefully controlled and empirically designed environments.

The progression of interviews transpired as a resilient mechanism for engagement, facilitating semi-structured dialogues with prominent figures in the field of artificial intelligence. The utilization of qualitative inquiry principles in these dialogues facilitated the exploration and expression of personal and professional perspectives on the ethical, social, and legal aspects of artificial intelligence. The subsequent interview transcripts served as valuable sources of detailed narratives, which were then analyzed to identify common themes. These themes provided additional support, contrasting viewpoints, and enriched perspectives to the conclusions drawn from the literature review and content analysis.

3.1.2 Utilized Data Analysis Methods

In the pursuit of data sources that align with the fundamental nature of the study, non-probability sampling methodologies have gained prominence. Considering the unique characteristics of the research landscape and the need to access the knowledge of experts in artificial intelligence, purposive sampling emerged as the preferred methodology. The intentional method of participant selection allowed for the inclusion of individuals who possess extensive knowledge and experience in the field of artificial intelligence. This contributed valuable firsthand accounts, personal experiences, and cognitive insights to enhance the research process.

The application of the interpretive framework of content analysis was skillfully utilized to analyze and interpret the qualitative data obtained from a range of literary sources. The methodology employed in this study involved a rigorous process of categorizing and analyzing textual data. This approach facilitated the identification of recurring themes, patterns, and initial insights that are relevant to ethical considerations, societal impacts, and the complex legal framework that AI is intertwined with. The process of content analysis has the ability to transform data into narratives that are rich in knowledge, thereby providing insights into the various dimensions associated with artificial intelligence.

Data from the interviews were analyzed using theme analysis. Interviews with professionals, stakeholders, and people impacted by artificial intelligence (AI) technology were undertaken. Through thorough coding of interview transcripts and thematic analysis, recurrent themes pertaining to the social, legal, and ethical implications of AI were found. The obstacles and ramifications of AI incorporation into society were then thoroughly understood by interpreting these themes. Thematic analysis enhanced the research findings and contributed to a comprehensive investigation of the thesis issue by revealing complex viewpoints and insights from respondents.

In conclusion, the employed research methodologies have facilitated a comprehensive comprehension of the complex ethical, social, and legal framework created by artificial intelligence technologies. The utilization of qualitative research methods, extensive literature analysis, content analysis, and strategic interviews resulted in an enhanced understanding of the various complexities inherent in the ethical, social, and legal dimensions of artificial intelligence.

3.2 The Rationale for the Selection of the Methodology

The incorporation of a thorough literature review and the extraction of secondary data are essential in order to develop a robust foundational comprehension of the intricate terrain encompassing the ethical, social, and legal dimensions of artificial intelligence. The adoption of this particular methodology enables the integration of current knowledge, identification of crucial themes, and recognition of emerging trends within the field of artificial intelligence. The extensive array of academic sources provides valuable insights and diverse perspectives from experts and scholars, facilitating a comprehensive examination of the research issue within its appropriate context. Through an in-depth examination of existing research and pertinent documents, this study attains a comprehensive viewpoint that establishes a standard for interpreting and assessing the empirical findings.

The duality of data collection methodologies, encompassing both descriptive and experimental approaches, arises from the necessity to comprehend both broad patterns and specific operational intricacies of artificial intelligence (AI) technologies. The descriptive approach aims to comprehensively explore the ethical, social, and legal considerations, in line with the study's objective of providing a comprehensive perspective on the impact of artificial intelligence. In contrast, the utilization of experimental methodology provides the chance to investigate particular scenarios and use cases, thus facilitating a comprehensive examination of the implications of artificial intelligence within controlled contexts. The inclusion of various methodologies in this study allows for a comprehensive examination of the subject matter, while also uncovering subtle details that may be disregarded in more general analyses.

The utilization of non-probability purposive sampling is a deliberate strategic decision made in light of the research topic's specialized and domain-specific characteristics. The emphasis on involving AI experts and stakeholders is in line with the objective of accessing firsthand perspectives and insights from individuals who possess extensive experience in AI development, policy creation, and implementation. The utilization of this specific sampling technique guarantees that the research is enhanced with genuine narratives and experiences that form the foundation of the ethical, social, and legal intricacies linked to artificial intelligence. By directing attention towards this specific group of participants, the study seeks to uncover intricate viewpoints and perspectives that would otherwise remain unexamined in more comprehensive random sampling approaches.

The use of content analysis has proven to be a valuable method for extracting significance from a wide range of textual data collected from various sources. The selection of this methodology is consistent with the study's aim to comprehensively classify, encode, and analyze written material in order to identify recurring themes, patterns, and emerging insights. The study utilizes content analysis as a methodological approach to systematically extract and integrate significant information pertaining to ethical considerations, societal impacts, and legal dimensions of artificial intelligence (AI). The chosen methodology enables the research to effectively analyze the qualitative data by incorporating various perspectives into a cohesive framework.

The conduction of interviews as a method of data analysis is based on the necessity to obtain primary and personalized perspectives from individuals who are actively involved in the field of artificial intelligence. This particular methodological approach facilitates an in-depth exploration of the ethical, social, and legal aspects, providing a platform for participants to openly discuss their experiences, beliefs, and expert viewpoints. Interviews play a crucial role in connecting theoretical frameworks with real-life experiences, enhancing the qualitative findings obtained from literature reviews and content analysis. The interactive nature of interviews fosters a conducive environment for participants to engage in thoughtful introspection regarding the intricate facets that contribute to a comprehensive understanding of the multifaceted challenges associated with artificial intelligence.

Through the incorporation of various research methodologies, this study enhances its comprehension of the ethical, social, and legal challenges of artificial intelligence (AI), thereby guaranteeing a thorough and knowledgeable examination of the research issue.

3.3 Empirical Investigation: Interviews

The conduction of interviews has been instrumental in facilitating comprehensive data gathering for the thesis, which explores the intricate terrain encompassing the social, legal, and ethical dilemmas presented by artificial intelligence. By actively participating in extensive discussions with experts, stakeholders, and practitioners in the field of artificial intelligence (AI), I have been fortunate to acquire invaluable insights and perspectives that enhance the qualitative depth of the research.

Through these interviews, an opportunity to delve into the nuanced dimensions of artificial intelligence's influence on society, the complexities of legal structures, and the ethical quandaries confronted by developers and users of AI has been utilized. The inclusion of firsthand narratives and experiences shared by the interviewees has contributed a humanistic perspective to the intricate matters being examined. The participant’s sincere and forthright answers have provided valuable insights into the practical obstacles and ethical quandaries that arise in relation to artificial intelligence (AI). This has contributed to a comprehensive comprehension of the various ways in which AI is perceived, encountered, and governed across diverse settings.

In conclusion, the inclusion of interviews in my data collection methodology has not only bolstered the validity and comprehensiveness of my research, but has also cultivated a heightened understanding of the practical ramifications of the widespread integration of artificial intelligence in our society.

The interview has been conducted with a total of five candidates, comprising:

1- The first Interviewee is a researcher specializing in deep learning and an AI developer employed at Linkthat. The individual in question has held the position of Machine Learning Engineer at Canva and is presently engaged as an AI researcher and developer with the objective of creating innovative AI solutions.

2- The Interviewee is a machine learning engineer plays a crucial role in assisting organizations in the development and implementation of artificial intelligence (AI) solutions to address practical challenges encountered in real-world scenarios. The individual possesses extensive expertise in the field of AI training and is presently employed as an AI data scientist in Mexico.

3- The interviewee holds the position of a healthcare industry architect at Microsoft, where his primary responsibility entails assisting healthcare customers and partners in their transition towards utilizing the intelligent cloud through Microsoft Azure.

4- The interviewee is a Cloud Solution Architect at Microsoft, specializing in providing support for Mission Critical systems (SfMC) with the aim of ensuring comprehensive care for such systems.

The interview encompassed a series of inquiries, which were as follows:

- What are the primary ethical dilemmas that arise in relation to the advancement and implementation of AI technologies?
- How can AI developers effectively establish transparency and accountability within their algorithms in order to effectively address ethical concerns?
- How has AI influenced or disrupted traditional job roles and industries in your experience?
- Could you provide an analysis of the various instances in which artificial intelligence (AI) has exerted a substantial influence on human behavior or social dynamics?
- In the context of your professional background, have you encountered any legal complexities or uncertainties pertaining to artificial intelligence (AI) technologies? Could you kindly provide some examples, please?
- In your perspective, what are the most pressing domains necessitating additional research or policy formulation to effectively navigate the intricacies of artificial intelligence?
- What recommendations can be provided to policymakers, AI developers, and stakeholders in order to more effectively tackle the ethical, social, and legal challenges presented by AI?

4 Implementation / Testing / Evaluation Results

4.1 Interview Perceptions and Interpretations

This section discusses the perspectives and suggestions provided by experts in the field of artificial intelligence (AI) pertaining to the ethical, social, and legal implications of AI on society. Additionally, it explores strategies to develop AI systems that are reliable and safeguarded. The comprehensive responses are provided in Appendix B.

Interviewee 1:

When asked about suggestions for tackling the complex issues associated with artificial intelligence, the AI researcher interviewed shared valuable perspectives. The speaker placed significant emphasis on the significance of education and issued a warning against yielding to exaggerated enthusiasm surrounding artificial intelligence (AI), urging individuals with vested interests to acquire a comprehensive understanding of the practical ramifications of this technology. Furthermore, he emphasized the importance of dedicating time to understanding artificial intelligence (AI) systems, with a particular focus on the crucial considerations of robustness and security. The researcher advocated for the need of collaborative endeavours between academia, industry, and government in order to foster the development of sophisticated security measures. The researcher underscored the importance of safeguarding against vulnerabilities.

Furthermore, the researcher expressed apprehensions regarding the potential of AI to foster a perception of indolence among its users. The increasing reliance on AI systems for task handling and immediate problem-solving may potentially diminish individual’s inclination towards critical thinking and problem-solving, thereby impacting their work habits and productivity. When prompted to present examples of significant effects of AI on human behaviour and social dynamics, the AI researcher provided insightful perspectives on various crucial domains. Initially, the author acknowledged the psychological and social ramifications of artificial intelligence (AI), highlighting the growing dependence of individuals on AI platforms such as ChatGPT. The increasing reliance on artificial intelligence

(AI) for information retrieval and human-computer interaction possesses the capacity to fundamentally alter individual’s perceptions and interactions with technology, thereby exerting an influence on their behaviours and attitudes.

Moreover, the researcher emphasized a notable alteration in decision-making procedures, whereby a considerable number of individuals currently depend on artificial intelligence (AI) as a source of guidance when making decisions. The alteration in the dynamics of decision-making, which is influenced by recommendations and predictions generated by artificial intelligence, carries significant implications for personal autonomy and individual agency.

In relation to areas necessitating additional investigation and policy formulation in the field of artificial intelligence (AI), the researcher has identified ethical considerations in AI, security concerns, environmental sustainability (also known as green AI), and the optimization of efficiency as pivotal domains. The speaker underscored the ethical obligation to develop artificial intelligence (AI) systems devoid of bias and stressed the necessity of implementing robust security protocols. In addition, he espoused the adoption of environmentally sustainable AI technologies and resource-efficient models as means to address and alleviate the negative effects on the environment.

The researcher emphasized the importance of meticulous data selection, thorough model training, rigorous testing, and continuous quality checks as crucial measures to ensure ethical development of AI algorithms with regards to transparency and accountability. Regarding occurrences of bias and discrimination in artificial intelligence (AI), the researcher emphasized specific situations such as loan approval processes and job application evaluations. He emphasized the significance of thorough evaluation, the reduction of bias, and continuous monitoring from an ethical standpoint.

Ultimately, the researcher presented concrete illustrations of the effects of artificial intelligence (AI) on job positions and sectors in the real world. These examples encompassed the alteration of design responsibilities through the utilization of AI metric tools, as well as the impact of AI on the domains of copywriting and coding. Additionally, the author observed the involvement of artificial intelligence (AI) in problem decomposition, wherein AI generates code for smaller components within larger systems. This highlights the significant impact of AI across different industries and emphasizes the necessity for adaptation. Taken together, these observations underscore the significance of well-informed and prudent artificial intelligence (AI) advancement, the key domains necessitating additional investigation and policy formulation, and the ethical deliberations crucial for the responsible implementation and incorporation of AI across various sectors.

Interviewee 2:

During the interview, the AI researcher provided insights into various critical challenges and ethical considerations associated with the domain of machine learning and artificial intelligence (AI). One of the primary challenges that was discussed pertained to the application of user data for the purpose of training machine learning models. The researcher highlighted the necessity of tackling this challenge through the inclusion of terms and services clauses, emphasizing the significance of protecting personal information and upholding privacy considerations.

Another notable challenge that was emphasized is the existence of biases within the data utilized for training artificial intelligence (AI) models. The researcher presented a specific instance to demonstrate how biases present in data, such as the association between gender and salary, can unintentionally sustain discriminatory practices if not adequately addressed during the training phase. This highlights the ethical obligation to acknowledge and address biases in order to guarantee that AI systems deliver equitable and impartial predictions.

The interviewee expressed apprehensions regarding the automation of jobs and the potential ramifications for the labour force as artificial intelligence (AI) progresses further. The researcher emphasized the necessity of proactively addressing the challenge of job displacement as the field of artificial intelligence (AI) continues to advance, while also acknowledging that it may not be an immediate concern. Within the domain of education, the researcher examined the profound influence of artificial intelligence (AI), specifically focusing on Large Language Models (LLMs). The rapid production of high-quality essays by students poses a challenge to conventional pedagogical approaches. The researcher emphasized the importance for educational institutions to carefully contemplate the integration of AI technologies in order to augment education while safeguarding the development of student’s critical thinking and creative capacities.

Finally, the interviewee discussed the matter of content recommendation on social media platforms. He emphasized the tendency of AI-powered content recommendation algorithms to prioritize user engagement at the expense of content quality and accuracy. This prioritization has the potential to facilitate the spread of false or misleading information and further reinforce user biases. This observation necessitates a reassessment of content recommendation algorithms in order to ensure their alignment with wider ethical goals.

In general, the observations derived from the interview highlight the complex difficulties and ethical deliberations that are inherent in the fields of artificial intelligence (AI) and machine learning. These challenges encompass a wide range of issues, including safeguarding data privacy, mitigating bias, and understanding the repercussions on employment, education, and the dissemination of content on social platforms. These considerations play a crucial role in providing guidance for the responsible development and implementation of AI technologies in the present-day society.

Interviewee 3:

During the interview conducted with an expert in the field of artificial intelligence (AI) in healthcare at Microsoft, the discussion centred around six key principles that are deemed essential for ensuring responsible and ethical AI practices. These pillars, including fairness, reliability, safety, privacy, inclusiveness, apparency, and accountability, were emphasized as crucial components in the development and deployment of AI technologies. The significance of transparency in the development of artificial intelligence (AI) was underscored by the expert, who urged developers to openly communicate their data sources and exhibit their commitment to data profiling, quality assessment, and distribution metrics. The importance of mitigating bias and promoting diversity in datasets was emphasized as a means of preventing the development of biased artificial intelligence algorithms. The interviewee emphasized the importance of remaining vigilant regarding edge circumstances and the necessity of incorporating disclaimers when employing AI in key domains such as medical diagnostics. The topic of drift detection in artificial intelligence (AI) models was examined, with a particular focus on the significance of continuously monitoring the performance of models in real-world situations and adjusting them to accommodate changing populations. In relation to the influence of artificial intelligence (AI) on occupational positions, the specialist elucidated that AI ought to be perceived as a collaborative partner, augmenting human capacities rather than supplanting them. Regarding the future, the interviewee conveyed apprehensions regarding the ethical aspects of AI prediction models in domains such as healthcare. They underscored the imperative for comprehensive investigation, policy formulation, regulatory measures, and establishment of societal norms to effectively tackle these difficulties. This interview offers significant insights into the responsible advancement and ethical considerations of artificial intelligence (AI), specifically within the realm of healthcare. It also prompts relevant inquiries for subsequent research endeavours and the formulation of policies.

Interviewee 4:

During an interview with a security specialist from Microsoft, a comprehensive examination of the fundamental elements pertaining to the ethical considerations of artificial intelligence (AI) and its ramifications on the industry were deliberated. One prominent ethical concern that was emphasized is the matter of privacy, wherein Microsoft takes an active approach by incorporating privacy as a set of guidelines to assist developers in adhering to ethical issues right from the initial phases of the design process. Furthermore, the issue of accountability has surfaced as a prominent consideration, particularly within the realm of security. In response to the inquiry regarding strategies for AI developers to foster transparency and accountability, the interviewee acknowledged the nascent nature of rules in this domain and underscored the significance of toolkits as valuable resources to aid developers in effectively addressing ethical dilemmas.

The expert highlighted notable effects of AI on conventional employment positions and industries, with a specific emphasis on its influence on security operations, particularly in the domain of incident management. The advent of artificial intelligence (AI) has significantly transformed the security domain by facilitating the provision of comprehensive insights into various incidents. This technological advancement has revolutionized security practices, allowing for the use of tactics such as reverse engineering attacks. Consequently, AI has empowered security professionals to effectively identify the underlying origins of breaches.

The interviewee emphasized the significant impact of AI on human behaviour and social dynamics, specifically highlighting the efficacy of ChatGPT and the incorporation of AI functionalities into Microsoft products as illustrative instances. Additionally, the expert emphasized the necessity for ongoing research in the domains of security breaches targeting machine learning systems and the wider societal implications of artificial intelligence, highlighting the significance of future exploration into these subjects especially when it comes to the cyber-attacks on the algorithms and the impact of AI on human behaviour.

In brief, this interview with the security expert from Microsoft provides insights into the ethical quandaries associated with artificial intelligence (AI), the actions being implemented to tackle them, and the profound impacts of AI on employment positions, sectors, human conduct, and security. It underscores the persistent requirement for research and attentiveness in these domains.

4.1.1 Interviews Analysis of Results

The table below demonstrates the final results of the interviews in the form of challenges, new findings and similarities amongst the interviewees.

Interviewee	Challenges	Findings	Similarities
1	• lack of educations regarding AI • Robustness and security • Dependence and indolence	• Effect of AI on Environment • Real world effect of AI on jobs • AI impact on human psychology	• Ethical AI Development • Data Privacy and Bias Mitigation • Job displacement concerns • Transparency and Accountability • Educational and Content Concern
2	• User data protection • Content recommendation algorithms • Educational impact of AI	• Content Quality vs. Engagement	
3	• Concerns about Ethical AI in Healthcare	• Bias Prevention Through Diversity • AI as a Collaborative Partner • Principles for Responsible AI	
4	• Privacy and Accountability	• Transparency and Toolkit Usage • Transformation of Security Operations	

Table 1 Comparison of Interview Findings

In order to recognize the recurrent themes that surfaced from the interviews conducted with professionals in the domain of artificial intelligence (AI). The interviews consistently highlighted five critical points. The crucial ethical aspects of data privacy and the reduction of bias in AI systems were emphasized, emphasizing the importance of safeguarding user data and promoting justice in the development of AI. Furthermore, the interviewees consistently highlighted the importance of ethical progress in the field of artificial intelligence (AI), underscoring the necessity for clear norms and a steadfast dedication to accountability. Furthermore, the interviews recognized issues pertaining to job displacement caused by artificial intelligence (AI) and its ramifications for the labour market, underscoring the necessity of taking proactive measures to tackle this issue. Furthermore, there was a persistent emphasis on the significance of transparency and accountability in the development of artificial intelligence (AI). This encompassed various issues including effective sharing of data sources, continuous monitoring of models, and the responsibility of developers. Finally, the interviews explored the implications of artificial intelligence (AI) on education and the usage of content recommendation algorithms in social media. These discussions shed light on the intricate process of incorporating AI into education while simultaneously safeguarding critical thinking skills and maintaining a delicate equilibrium between content quality and user engagement. The identification of these recurring themes offers significant contributions to the understanding of ethical and practical implications within the domain of artificial intelligence, Discussion

This analysis supports the theory that, artificial intelligence (AI) possesses the capacity to profoundly transform numerous sectors, yet it concurrently introduces ethical, social, and legal complexities. Additional research and the implementation of legislative measures are necessary to guarantee a fair and accountable integration of artificial intelligence (AI) into society.

4.2 Discussion and Interpretation

The results indicate that Artificial Intelligence (AI) has brought about significant transformations across diverse sectors, encompassing healthcare, humanitarian aid, and social activities, with the capacity to facilitate economic growth, societal advancement, and human well-being. Nevertheless, the rapid growth of this phenomenon gives rise to a range of ethical, legal, and social quandaries that necessitate continuous examination. The restricted level of understandability exhibited by AI-driven technology, apprehensions regarding the safeguarding of data and privacy, as well as ethical deliberations, present notable hazards to users, developers, and governmental parties. The primary objective of this study was to examine the ethical, social, and moral ramifications associated with the implementation of artificial intelligence.

The incorporation of artificial intelligence (AI) into routine tasks has given rise to ethical considerations pertaining to autonomy, accountability, and bias. The intricacy of accountability is amplified when artificial intelligence (AI) systems depend on intricate algorithms, and there is a mounting concern regarding the existence of bias within AI systems. In order to achieve fair and unbiased decision-making, it is imperative to possess a high level of technical expertise and a thorough comprehension of social dynamics and ethical principles.

The data suggest that The Age of Artificial Intelligence (AI) represents a pivotal juncture in the annals of human civilization, as it is widely posited that the advent of intelligent machines and algorithms holds the potential to augment and amplify human capacities. The progress made in artificial intelligence (AI) technology has resulted in its utilization across various domains, including natural language processing, computer vision, machine learning, and robotics. These advancements have proven advantageous for sectors such as healthcare, finance, and transportation,etc. Nevertheless, it is imperative to acknowledge that artificial intelligence (AI) has the potential for misuse or the manifestation of detrimental behaviours. Consequently, the significance of law, ethics, and technology in governing AI systems is progressively paramount. According to the studies, Artificial Intelligence (AI) can be categorized into two main types: weak AI and strong AI. Weak AI refers to AI systems that demonstrate intelligence in specific tasks or domains, while strong AI aims to replicate the cognitive abilities and characteristics of a human mind. The distinction between weak and strong artificial intelligence (AI) was first introduced by philosopher John Searle in

1980, thereby initiating a critical examination of the fundamental characteristics of AI. The concept of Strong AI necessitates a level of generality, as opposed to general AIs that lack the capability to exhibit true cognitive abilities but can simulate general intelligence.

Scholars from diverse fields are collaborating to advance the progress of creating resilient artificial general intelligence (AGI), capable of performing a broad spectrum of tasks at a level of intelligence comparable to humans. In contrast to narrow artificial intelligence (AI), which exhibits limitations in performing specific tasks. The notion of self-improving general-purpose systems pertains to the capacity of a system to augment its functionalities and adjust to artificial intelligence. These results build on existing evidence of the proliferation of disinformation, which refers to the deliberate dissemination of false or biased information with the intention to mislead, is an increasingly prevalent phenomenon that presents significant challenges to both commercial enterprises and democratic systems. Starting with Artificial intelligence (AI) agents which possess the capacity to produce textual content that supports various perspectives, enabling them to potentially influence public discourse by disseminating disinformation and promoting fabricated agendas. Corporations have the capacity to employ artificial intelligence (AI) agents in order to create narratives that appear credible in relation to their corporate social responsibility (CSR) initiatives. This utilization of AI agents can potentially facilitate the process of greenwashing by augmenting it with machine-based practices. Adversarial stakeholders have the capability to employ artificial intelligence agents in order to produce plausible narratives that portray corporations in an unfavourable manner.

According to the recent scientific implication, In order to optimize the utilization of AI agents, it is imperative to establish and incorporate a concept of expertise and prestige pertaining to the source within the learning procedure. The revaluation of the concept of expertise within AI agents is of utmost importance, as it is based on the statistical assumption that a substantial and diverse collection of sources will effectively encompass the knowledge pertaining to a given subject matter.

Artificial Intelligence (AI) that is commonly used nowadays is known as Narrow AI or Weak AI, it demonstrates efficacy within a particular domain. However, its performance is heavily reliant on both training data and programming, which are closely intertwined with the concepts of big data and individual contributions. The development of artificial intelligence (AI) systems is highly dependent on extensive quantities of data, encompassing private and personal information, necessitating appropriate measures to ensure protection against unauthorized utilization and exploitation. The opacity of AI systems, specifically those employing machine learning and deep learning techniques, can hinder thorough examination, thereby increasing the potential for malevolent exploitation.

It was clearly demonstrated that AI has the potential to inherit human biases, including those related to gender and race, which constitutes a significant factor to consider. Artificial intelligence (AI) systems continue to undergo training facilitated by human operators and rely on datasets generated by humans. These datasets are subsequently assimilated by AI systems and subsequently employed in practical, real-world scenarios. The consideration of accountability arises when an AI system fails to successfully execute a designated task. The Resolution of the European Parliament urges the prompt establishment of a legislative framework that governs robots and artificial intelligence (AI) systems with the ability to anticipate and adapt to scientific advancements projected in the near future. The prevalence of selection bias in datasets utilized for the development of AI algorithms is a widespread issue. This bias is evident in automated facial recognition systems and the datasets associated with them, leading to diminished accuracy when identifying individuals with darker skin tones, especially women.

It is imperative for organizations to adhere to ethical standards during the development of artificial intelligence (AI) systems. The process of shifting from curation to creation encounters obstacles in the efforts to address biases and uphold fairness in AI-generated content. Artificial intelligence (AI) systems have the ability to acquire knowledge from large datasets, which may contain biases that can be reproduced in the generated content. This can result in the production of "real-time" deepfakes. The utilization of deepfakes has the potential to manipulate electoral processes, leading to the dissemination of misleading information among voters and ultimately eroding the foundations of democratic systems. Nevertheless, the utilization of deepfake technology also engenders apprehensions regarding its prospective application in military operations, wherein the dissemination of deceptive information and the fabrication of evidence may heighten hostilities and exacerbate armed conflicts.

The data contribute a clearer understanding of how Artificial intelligence (AI) has brought about a significant transformation in various industries through its ability to replicate human cognitive functions. This has resulted in the optimization of operational processes and a profound impact on human interactions. Nevertheless, the legal framework encounters various obstacles in effectively regulating these systems, encompassing issues related to liability, responsibility, and accountability. With the increasing autonomy of AI systems, it is imperative to establish comprehensive guidelines for developers, users, and stakeholders. The societal ramifications of AI's extensive integration encompass employment displacement and the potential amplification of societal disparities as a result of biased algorithms and decision-making driven by data. The imperative for ethical and responsible deployment of artificial intelligence (AI) is of paramount importance, given the growing autonomy and independence of AI systems.

According to many studies, The importance of safeguarding data privacy and protection is paramount within the realm of Artificial Intelligence (AI) and its application in daily existence. The utilization of artificial intelligence (AI) engenders apprehensions pertaining to data accessibility, the categories of information it can retrieve, and the possibility of privacy infringements. The capacity of artificial intelligence (AI) to recognize patterns, acquire knowledge, and make predictions pertaining to individuals and collectives is substantial; however, it has the potential to generate data that surpasses what an individual intentionally divulged. The utilization of predictive technologies has the potential to deduce personal information pertaining to individuals who have made a deliberate decision to withhold such information. This gives rise to inquiries regarding the definition of personal information and its applicability to principles governing the privacy of information. Artificial intelligence (AI) methodologies, specifically machine learning, heavily depend on extensive datasets for the purpose of training and evaluating algorithms. This reliance on large amounts of data can potentially conflict with the principle of limited data collection. The proliferation of Internet of Things (IoT) devices, smartphones, and web tracking has resulted in the accumulation of data from these devices, which is subsequently utilized as input data for artificial intelligence (AI) systems.

The rapid analysis of extensive datasets, including personal information, by artificial intelligence (AI) holds potential advantages in enhancing information accessibility. Nevertheless, the improper utilization of social media platforms can result in significant repercussions, including the generation of financial gains for these platforms and the manipulation of consumer actions. AI-powered information systems have the potential to erode democratic processes through the erosion of trust, the exacerbation of social divisions, and the incitement of conflicts, ultimately leading to adverse public consequences. AI-driven surveillance can be employed by governments and influential entities to exert direct control and suppress individuals.

Furthermore, Artificial intelligence (AI) possesses the capacity to bring about a transformative impact on military and defence systems. However, it also presents a noteworthy peril in terms of depersonalizing lethal force and engendering the development of Lethal Autonomous Weapon Systems (LAWS). Autonomous weapons possess the capability to independently detect, identify, and engage human targets without the need for human intervention, thereby signifying a significant paradigm shift in the realm of warfare. These objects have the potential to be manufactured in large quantities at a low cost and can be readily deployed to cause harm on a large, industrialized level. There has been a continuous discourse surrounding the prevention of the proliferation of lethal autonomous weapons systems (LAWS) and the imperative to safeguard them against cyber infiltration or misuse. Artificial intelligence (AI) possesses the capacity to adversely affect the overall welfare of a substantial portion of society by means of social determinants of health, including exerting control and manipulation over individuals, as well as the potential for causing extinction. Furthermore, it is plausible that artificial intelligence (AI) may contribute to the continuation of moral shortcomings and hinder the progress of moral values. The evaluation of the consciousness and level of consciousness in highly developed artificial intelligence (AI) presents considerable difficulties. However, neglecting to prioritize the welfare of these AI entities could potentially result in severe risks to our survival.

According to many scholars, it was mentioned that the utilization of AI technology gives rise to apprehensions regarding the possibility of bias and discrimination. Bias can arise due to inaccurate presumptions or data that does not adequately represent the population, resulting in prejudiced judgments influenced by variables such as race, gender, or socioeconomic standing. Artificial intelligence (AI) systems possess a level of impartiality that is directly proportional to the impartiality of the data on which they are trained. Consequently, if the training data exhibits bias, the resulting AI system will inevitably reflect that bias. In order to mitigate bias, it is imperative that AI systems undergo training using a diverse array of data and undergo regular audits. Three primary sources of bias commonly manifest in artificial intelligence (AI) systems: data bias, algorithmic bias, and user bias.

This analysis supports the theory that Artificial intelligence (AI) systems are progressively enhancing their capacity to execute tasks that have conventionally been carried out by human beings. This advancement has the potential to result in job displacement, economic upheaval, and the necessity to provide individuals with retraining opportunities. The matter at hand is intricately connected to apprehensions regarding privacy, as the implementation of AI technology has the potential to heighten the financial vulnerability of workers and compel them to relinquish their privacy. The increasing complexity of AI-related occupations has led to projections indicating that the implementation of AI-driven automation will result in a significant number of job displacements, ranging from tens to hundreds of millions, over the course of the next ten years. The ramifications will be particularly notable in low and middle-income nations, as it will entail the displacement of employment opportunities that necessitate lower levels of skill. Nevertheless, it is crucial to acknowledge that the planet's ability to sustain the exploitation for the purpose of economic production is limited, and ensuring a fair distribution of productivity within society is not a guaranteed outcome. The advent of automation has predominantly led to a reallocation of income and wealth from labour to capital proprietors, thereby intensifying global disparities in wealth. The current state of affairs regarding the scarcity or obsolescence of employment in society is marked by a notable deficiency in comprehending the psychological and emotional ramifications. Furthermore, there appears to be a dearth of attention given to the formulation of policies aimed at mitigating the adverse health consequences associated with unemployment.

Moreover, throughout the study, it was noticable that Cyber threats pose a significant challenge for security personnel, causing significant harm to networks. Advanced technologies, such as artificial intelligence (AI), help enhance operational efficiency and protect organizations against cyber threats. AI can efficiently manage and process large volumes of data, enabling data transmission and detecting hidden hazards. However, AI can also lead to duplicative processes, reducing efficiency. AI can handle repetitive cyber security tasks, alleviating the burden on security personnel while emulating human characteristics and excluding inherent limitations. This helps identify and mitigate inherent security vulnerabilities on a recurring basis, ensuring the security of the network.

The study demonstrates a correlation between Trust in artificial intelligence (AI) and the social,ethical and legal impact it has as trust is crucial for AI‘s success, as it fosters user acceptance, promotes transparency and accountability, and mitigates AI bias. Trustworthy AI systems can offer justifications for their decisions, ensuring transparency and accountability. The concept of trust is also linked to the ethical advancement of AI, emphasizing user well-being, privacy, and adherence to ethical guidelines. The High-Level Expert Group on Artificial Intelligence (AI HLEG) endorses a comprehensive structure for Trustworthy AI, emphasizing three fundamental elements: adhering to laws and regulations, embodying ethical principles and values, and possessing robustness. Each component must function synergistically and exhibit overlapping functionality, and society should strive to harmonize them if conflicts arise. The successful integration of AI into society depends on trust, as it contributes to user acceptance, promotes fairness, and upholds ethical standards.

The guidelines by the European Union is proven to provide a comprehensive framework for trustworthiness in AI systems, focusing on ethical principles and practical application within socio-technical systems. They aim to foster and secure ethical and robust AI systems, addressing legality and potential opportunities and critical concerns. The guidelines are intended for all stakeholders

AI ethics are becoming increasingly important for governance, but existing mechanisms often fail to ensure ethical alignment. Organizations like Google, Intel, Oxford, Cambridge, and Stanford have developed ethical guidelines for AI systems, but implementation remains discretionary. The industry lacks tools and incentives to translate ethics principles into verifiable and actionable criteria. Ethics-based auditing, involving independent auditors to evaluate safety, security, privacy, and fairness assertions, is a promising approach. Professional services firms like Microsoft and Deloitte are developing frameworks to assist clients in creating and implementing AI systems. However, it is crucial to maintain a realistic perspective on the achievable outcomes of ethics-based auditing. It plays a significant role in promoting and upholding good governance principles, similar to how businesses rely on physical infrastructures for success.

These results should be taken into account when considering how to achieve Trustworthy AI, inclusivity and diversity are crucial throughout the system's life cycle. This includes engaging stakeholders, ensuring equitable access, and removing inadvertent biases. Data sets used by AI systems may contain historical biases, incompleteness, and bad governance models. Addressing these issues through oversight processes and hiring from diverse backgrounds can prevent prejudice and discrimination. Encouraging fair competition and hiring from diverse backgrounds can also help ensure diversity in AI systems.

When it comes to suggested mitigations and precautions to be taken to ensure a safe and trustworthy AI, many scholars have came to an agreement that AI model‘s security involves implementing advanced algorithms like AES-256 for data encryption, secure storage, and secure transmission. Access controls are also crucial to restrict authorized individuals from accessing confidential information. An informed user base is essential for defence against security threats, including phishing attacks and deepfake manipulation. Regular training sessions can help users recognize potential security threats and understand appropriate reporting protocols. Principles governing secure AI design include data minimization, privacy incorporation, secure data handling protocols, user consent, anonymization, and privacy-preserving technologies. Software security architecture is crucial for protecting data, models, and algorithms from unauthorized access, manipulation, or misuse. A thorough analysis of potential risks is necessary, and a comprehensive AI security architecture and deployment solution should be developed. Acquiring and enhancing proficiency in software security architecture for AI systems involves identifying security requirements and adhering to regulations, standards, and best practices. Furthermore, selecting appropriate security patterns for your AI system, addressing common security issues in software design and implementation is necessary for ensuring safety and security of algorithms. Addressing knowledge gaps among team members is crucial in the interdisciplinary nature of machine learning applications. The security design pattern involves segregating input from features, which are extracted from raw input for training models. However, preprocessing steps can cause issues in reproducibility during prediction, so it's essential to differentiate between inputs and features.

AI systems are assessed through a task-oriented methodology, which signifies their growing significance without necessarily implying a commensurate rise in cognitive capabilities. The domain of artificial intelligence (AI) exhibits a wide range of subfields and is experiencing rapid expansion, rendering its classification challenging. According to a study conducted by Russell and Norvig (2016), there is evidence to suggest that artificial intelligence (AI) is undergoing a shift from rule-based systems to machine learning techniques. This transition allows machines to acquire knowledge and enhance their performance through learning processes. The anticipated impact of this shift is poised to bring about significant transformation across multiple industries. Nonetheless, the absence of secure environments for educational and exploratory endeavors within public service may impede the capacity of representative governments to effectively address challenges. Furthermore, the utilization of AI in the realm of public service has the potential to be exploited for the purposes of surveillance and monitoring, thereby impinging upon the fundamental rights and societal entitlements of individuals.

The utilization of natural language processing and machine learning methodologies in platforms such as ChatGPT is significantly transforming consumer content by facilitating the development of dynamic and interactive experiences. The transition from human curation to AI-driven content creation presents advantages such as customized responses, efficient handling of extensive inquiries, and user preference customization.

4.2.1 Limitations of the Research

The present study has endeavored to conduct a thorough analysis of the ethical, legal, and social dilemmas presented by artificial intelligence (AI). However, it is crucial to recognize certain inherent constraints within this research. The primary objective of this study is to gain a comprehensive understanding of the ethical, social and legal dimensions of artificial intelligence (AI). Additionally, it briefly addresses security concerns, specifically adversarial attacks, but does not extensively explore advanced research on safeguarding machine learning algorithms against such attacks. Furthermore, it is widely recognized that AI has a psychological impact on human behavior. However, a comprehensive examination of this facet is imperative in order to gain a complete understanding of the extent to which AI influences human behavior. Furthermore, the primary focus of this study revolves around the regulations and policies pertaining to artificial intelligence (AI) specifically within the European context. It is important to note that this regional specificity may restrict the applicability of the research findings to a broader global context, as AI regulations and policies can exhibit substantial variations across different regions and countries. The ethical analysis in this study is inherently subjective and open to interpretation, as different individuals and groups may possess diverse ethical perspectives, thereby introducing subjectivity into the analysis. Ultimately, the research incorporates interviews as a method of data collection, which has the potential to introduce bias in the selection of participants, despite attempts to ensure a varied and representative sample. It is important to take into account these limitations when interpreting the findings and conclusions of this research. However, it is worth noting that the research provides valuable insights into the ethical, legal, and social challenges of artificial intelligence, thereby establishing a basis for future research efforts in these significant domains.

4.2.2 Recommendations for Future Research

Based on the comprehensive investigation undertaken in this study concerning the ethical, legal, and social complexities linked to artificial intelligence (AI), a number of significant recommendations arise from the knowledge acquired, incorporating valuable contributions from interviews. Initially, the participants emphasized the crucial necessity of conducting further research on 'green AI' and its environmental consequences. They proposed that additional studies in this field are required to enhance our comprehension of AI's ecological impact and develop strategies to reduce it. In addition, Future studies should take into account the importance of prioritizing the advancement of machine learning (ML) architectural patterns in order to improve their security and deployment, in line with the concerns expressed by the participants.

In order to enhance the robustness of artificial intelligence (AI) systems, it is imperative to cultivate collaborative efforts across academia, industry, and government sectors. This collaboration will facilitate the advancement of sophisticated security protocols to counter adversarial attacks. The issue of the psychological impact of artificial intelligence (AI) on human behavior, specifically in relation to user trust and societal consequences, necessitates a demand for more extensive scholarly investigation. Such research has the capacity to influence the development of AI systems that prioritize the mental well-being and welfare of users.

In order to effectively tackle the worldwide scope of challenges associated with artificial intelligence (AI), it is advisable to foster international cooperation and engage in meaningful discussions among policymakers. This collaborative approach aims to establish a comprehensive regulatory framework that surpasses geographical limitations, thereby facilitating the responsible and fair utilization of AI technologies. Simultaneously, it is imperative to provide support for initiatives that seek to increase public awareness and promote the development of digital literacy with regards to the capabilities and limitations of artificial intelligence. This will enable individuals to make well-informed decisions. Moreover, Further research is needed to establish interdisciplinary cooperation among professionals hailing from diverse domains, including artificial intelligence, ethics, law, sociology, and psychology, in order to effectively address the complex obstacles that emerge from the incorporation of AI into societal frameworks. Finally, it is crucial to actively engage with a wide range of stakeholders, including industry leaders, policymakers, researchers, and civil society organizations, in order to promote comprehensive solutions and guarantee the responsible development and implementation of artificial intelligence (AI). These recommendations collectively provide a comprehensive framework for addressing the ethical, legal, and social aspects of artificial intelligence (AI) and progressing towards a more ethical and inclusive AI ecosystem.

5 Conclusion

The advent of artificial intelligence has initiated a period of notable technological progress, leading to significant transformations across various aspects of our society. The widespread incorporation of Artificial Intelligence (AI) systems, algorithms, and machine learning models has had a profound impact on various aspects of our lives, professional endeavors, and societal engagements. As we traverse this dynamic and constantly changing terrain, it becomes increasingly apparent that the relentless advancement of artificial intelligence (AI) is accompanied by a myriad of complex challenges that encompass various realms such as ethics, society, and the legal system. This thesis undertakes an exploratory investigation to deconstruct, examine, and comprehend the complex challenges and their significant ramifications for both the current and future contexts.

Through my investigation of ethical obstacles, I have discovered the complex network of moral quandaries that is intertwined within the realm of artificial intelligence (AI) development and implementation. The emergence of decision-making systems powered by artificial intelligence has sparked a wide range of ethical inquiries. Transparency, accountability, and fairness have emerged as fundamental principles within the ethical dialogue pertaining to artificial intelligence. The rise of algorithms with inherent biases has emphasized the necessity for ethical principles in the advancement of artificial intelligence. The integration of artificial intelligence (AI) into our daily lives necessitates the prioritization of ethical considerations in its design, deployment, and governance.

The concept of ethical artificial intelligence (AI) extends beyond mere adherence to guidelines, encompassing a dedication to ethical values. The conscientious management of artificial intelligence's influence on society, economy, and humanity as a whole is required. The statement necessitates the acknowledgment that artificial intelligence (AI) should not be regarded solely as a tool, but rather as an entity with the capacity to fundamentally transform the very essence of our being. The ethical ramifications of artificial intelligence (AI) are extensive and wide-ranging, encompassing various aspects such as self-driving vehicles making moral judgments during accident situations and AI algorithms exerting influence on democratic processes. In order to ensure the alignment of AI systems with human values, it is imperative to incorporate the principles of transparency, fairness, and accountability at their core.

The societal impact of artificial intelligence (AI) is unquestionable. Artificial intelligence (AI) is causing significant disruptions in various sectors, leading to changes in employment patterns and giving rise to complex discussions regarding the ethical consequences of AI in social contexts. The legal framework pertaining to artificial intelligence (AI) is currently undergoing a dynamic phase of change, as it endeavors to keep pace with the swiftly advancing technological environment. Nevertheless, the existence of uncertainties remains, thereby requiring a unified methodology to tackle the legal ramifications of artificial intelligence. The implications discussed here span across various domains, including the privacy of personal data, and determining responsibility for decisions made by AI systems.

The proliferation of artificial intelligence (AI) has wide-ranging societal implications. The impact of technological advancements extends beyond mere technological progress, as it serves as a transformative force that is restructuring conventional job roles, modifying industries, and shaping human behavior. As artificial intelligence (AI) further integrates into our everyday routines, inquiries regarding the ethical implications of AI within social dynamics gain heightened relevance. The implications of artificial intelligence (AI) on employment, for instance, necessitate meticulous examination. Artificial intelligence (AI) possesses the capacity to enhance productivity and optimize various sectors; however, it also engenders apprehensions regarding the displacement of jobs. The ethical considerations pertaining to this matter can be categorized into two main aspects: firstly, the need to ensure fair allocation of advantages resulting from artificial intelligence (AI), and secondly, the necessity to minimize the socio-economic repercussions on individuals whose employment prospects may be jeopardized by automation.

The legal framework is also contending with the challenges presented by artificial intelligence. The issue of ownership and authorship is brought to the forefront when considering intellectual property rights in works generated by artificial intelligence. The distinction between content generated by humans and content generated by artificial intelligence is becoming progressively indistinct. The presence of this ambiguity underscores the need for a reassessment of copyright legislation in order to guarantee its continued applicability and equity in the era of artificial intelligence.

In addition, the preservation of data privacy emerges as a pivotal element within the legal complexities presented by artificial intelligence. The extensive data consumption of artificial intelligence has generated apprehensions regarding the safeguarding of personal privacy. Artificial intelligence (AI) exhibits the potential to extract valuable insights from extensive datasets; however, it also introduces vulnerabilities concerning data breaches and unauthorized access. The legal framework pertaining to data privacy, as illustrated by regulations such as the General Data Protection Regulation (GDPR), aims to achieve a harmonious equilibrium between fostering innovation and safeguarding individual rights. Nonetheless, the expeditious rate at which artificial intelligence (AI) is advancing necessitates the ongoing adjustment of these regulations to guarantee their efficacy and congruence with the ever-changing technological environment.

A comprehensive analysis of the obstacles presented by artificial intelligence necessitates an examination of the topic of security. There are numerous security concerns associated with artificial intelligence (AI), which include adversarial attacks, privacy breaches, and vulnerabilities within AI systems. The presence of adversarial attacks, which involve the deliberate manipulation of input data with the intention of deceiving AI algorithms, presents substantial risks in critical domains such as autonomous vehicles and medical diagnosis. The increasing processing of personal and sensitive information by AI systems has led to a growing concern regarding privacy breaches. The security of personal data in an AI-driven world is brought into question due to the capacity of AI to extract sensitive information, even from seemingly innocuous data.

As I explore the domain of AI security, it is imperative to also contemplate strategies for constructing AI systems that are robust in terms of security. The vulnerabilities and challenges that have been identified should not discourage us, but rather serve as motivation to develop comprehensive measures for protecting artificial intelligence. The development of secure artificial intelligence (AI) necessitates a comprehensive strategy that addresses various dimensions, including technical and ethical aspects.

One of the key aspects of addressing the security concerns associated with AI systems involves implementing technical measures to bolster their resilience against adversarial attacks. In order to address this matter effectively, it is imperative to employ methodologies such as robust machine learning and the creation of algorithms that possess inherent resistance to manipulation. Privacy-preserving artificial intelligence (AI) techniques play a crucial role in addressing privacy risks by allowing AI systems to process sensitive data while maintaining the confidentiality of the data. In addition, it is imperative to ensure the security of the AI supply chain, encompassing all stages from data collection to model deployment, in order to mitigate the potential exploitation of vulnerabilities by malicious entities throughout the AI development process.

The incorporation of ethical considerations is of equal importance in the development of secure artificial intelligence. In order to safeguard the core principles of transparency, fairness, and accountability, it is imperative that security measures are harmonized with ethical principles. The ethical frameworks that prioritize the well-being and rights of individuals should guide the development and deployment of artificial intelligence (AI). Constructing a robust artificial intelligence (AI) system encompasses more than just technical obstacles. It represents a moral obligation that necessitates the careful negotiation of the intricate convergence between technology and ethics.

Looking forward, the ethical, legal, and social dimensions of artificial intelligence (AI) continue to be dynamic and complex. The trajectory of artificial intelligence (AI) in the coming years is defined by a persistent process of development, characterized by progressions that will have profound effects on our societies and economies, the full extent of which we are only just starting to comprehend. As artificial intelligence (AI) technologies continue to advance and are utilized in various fields, it is essential for us, as individuals, researchers, policymakers, and global citizens, to adjust and address the forthcoming challenges and opportunities.

It is imperative that our actions in the development of artificial intelligence (AI) remain guided by ethical considerations. Ensuring transparency in the decision-making processes of artificial intelligence (AI) is of utmost importance. It is imperative that individuals and parties involved possess the capacity to comprehend and scrutinize the determinations rendered by artificial intelligence (AI) systems, particularly within vital sectors such as healthcare, finance, and criminal justice. In order to ensure that AI systems are held responsible for their actions, it is imperative to establish accountability mechanisms. Ensuring the integration of the principle of fairness is imperative in the development of artificial intelligence (AI) in order to mitigate the potential amplification of societal biases and discriminatory practices. In addition, it is imperative that the legal framework regulating artificial intelligence (AI) undergoes concurrent development with technological progress. The implementation of regulations specifically designed for artificial intelligence (AI) is crucial in order to effectively tackle the distinct challenges presented by this technology. The scope of these regulations should be comprehensive, covering a diverse range of concerns such as data privacy, intellectual property, and the attribution of liability in relation to decisions made by artificial intelligence systems. The involvement of international cooperation and agreements will be of utmost importance in influencing the global legal framework pertaining to artificial intelligence (AI).

7. References

[1] Müller, Vincent C. & Bostrom, Nick (2016). Future progress in artificial intelligence: A survey of expert opinion. In Vincent Müller (ed.), Fundamental Issues of Artificial Intelligence. Springer. pp. 553-571.

[2] Gelsinger, P., & Gelsinger, P. (2019). Data: the new science. Dell. https://www.dell.com/en-us/blog/data-the-new-science/ Accessed September 09, 2023

[3] Perc, M., Ozer, M. & Hojnik, J. Social and juristic challenges of artificial intelligence. *Palgrave Commun* **5**, 61 (2019)

[4] Tippins, Nancy T.; Oswald, Frederick L.; and McPhail, S. Morton (2021) "Scientific, Legal, and Ethical Concerns About AI-Based Personnel Selection Tools: A Call to Action," *Personnel Assessment and Decisions*: Number 7 : Iss. 2 , Article 1.

[5] Dash, Bibhu and Ansari, Meraj Farheen and Sharma, Pawankumar and Ali, Azad, Threats and Opportunities with AI-Based Cyber Security Intrusion Detection: A Review (September 2022). International Journal of Software Engineering & Applications (IJSEA), Vol.13, No.5, September 2022.

[6] Brendel, A.B.; Mirbabaie, M.; Lembcke, T.-B.; Hofeditz, L. Ethical Management of Artificial Intelligence. *Sustainability* **2021**, *13*, 1974.

[7] Cath C. 2018 Governing artificial intelligence: ethical, legal and technical opportunities and challenges.Phil. Trans. R. Soc. A 376: 20180080. http://dx.doi.org/10.1098/rsta.2018.0080

[8] Siau, Keng & Wang, Weiyu. (2018). Ethical and Moral Issues with AI.

[9] Wikipedia contributors. (2023). Algorithmic bias. Wikipedia. https://en.wikipedia.org/wiki/Algorithmic_bias#cite_note-Gillespie_et_al-16. Accessed October 08, 2023.

[10] Office of Victorian Information Commissioner. (2022, October 6). Artificial Intelligence and Privacy - Issues and Challenges - Office of the Victorian Information Commissioner. Office of the Victorian Information Commissioner.

https://ovic.vic.gov.au/privacy/resources-for-organisations/artificial-intelligence-and-privacy-issues-and-challenges/. Accesses August 07, 2023.

[11] Beware the privacy violations in artificial intelligence applications. (n.d.). ISACA. https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2021/beware-the-privacy-violations-in-artificial-intelligence-applications. Accessed August 07, 2023.

[12] McClelland, C., & McClelland, C. (2020). The impact of artificial intelligence - widespread job losses. IoT for All. https://www.iotforall.com/impact-of-artificial-intelligence-job-losses?ref=thedigitalspeaker.com. Accessed October 09, 2023.

[13] Artificial intelligence design must prioritize data privacy. (2022, May 20). World Economic Forum. https://www.weforum.org/agenda/2022/03/designing-artificial-intelligence-for-privacy/?ref=thedigitalspeaker.com. Accessed August 07, 2023.

[14] Van Rijmenam Csp, M. (2023). Privacy in the age of AI: Risks, challenges and solutions. Dr Mark Van Rijmenam, CSP | Strategic Futurist Speaker. https://www.thedigitalspeaker.com/privacy-age-ai-risks-challenges-solutions/#:~:text=The%20Issue%20of%20Data%20Abuse%20Practices&text=Additionally%2C%20AI%20can%20be%20used,can%20have%20serious%20privacy%20implications. Accessed August 14, 2023.

[15] AI and data privacy: protecting information in a new era. (February 26, 2023). Technology Magazine. https://technologymagazine.com/articles/ai-and-data-privacy-protecting-information-in-a-new-era. Accessed August 16, 2023.

[16] Belenguer, L. (2022b). AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI And Ethics, 2(4), 771–787. https://doi.org/10.1007/s43681-022-00138-8

[17] Polyakov, A. (2021, December 11). How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors). Medium. https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c. Accessed August 07,2023.

[18] Asif, M. (2022, January 6). Data Poisoning: When artificial intelligence and machine learning turn rouge. Medium. https://medium.com/analytics-vidhya/data-poisoning-when-artificial-intelligence-and-machine-learning-turn-rouge-d8038f423922. Accessed August 12, 2023.

[19] Dickson, B., & Dickson, B. (2021b, April 23). Machine learning: What are membership inference attacks? - TechTalks. TechTalks - Technology solving problems. and creating new ones. https://bdtechtalks.com/2021/04/23/machine-learning-membership-inference-attacks/. Accessed August 09, 2023.

[20] IEEE TCSP - IEEE Computer Security's technical community on security and privacy. (n.d.). https://www.ieee-security.org/. Accessed August 28, 2023.

[21] Van Bekkum, M., & Borgesius, F. Z. (2022). Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *ArXiv*. https://doi.org/10.1016/j.clsr.2022.105770

[22] Mökander, J., Floridi, L. Ethics-Based Auditing to Develop Trustworthy AI. *Minds & Machines* **31**, 323–327 (2021). https://doi.org/10.1007/s11023-021-09557-8

[23] Brendel, A.B.; Mirbabaie, M.; Lembcke, T.-B.; Hofeditz, L. Ethical Management of Artificial Intelligence. Sustainability 2021, 13, 1974. https://doi.org/10.3390/ su13041974

[24] Kieslich, Kimon & Starke, Christopher & Dosenovic, Pero & Keller, Birte & Marcinkowski, Frank. (2020). Artificial Intelligence and Discrimination. How does the German public think about the discrimination potential of artificial intelligence?.

[30] Salau, Ayodeji & Barud Demilie, Wubetu & Akindadelo, Adedeji & Nnenna, Eneh. (2022). Artificial Intelligence Technologies: Applications, Threats, and Future Opportunities.

[31] Chukhnov, A & Ivanov, Yuriy. (2021). Algorithms for Detecting and Preventing Attacks on Machine Learning Models in Cyber-Security Problems. Journal of Physics: Conference Series. 2096. 012099. 10.1088/1742-6596/2096/1/012099.

[32] A P Chukhnov and Y S Ivanov 2021 *J. Phys.: Conf. Ser.* **2096** 012099

[35] Jia, H. (2022, May 9). ML Model Security – Preventing The 6 Most Common Attacks - Excella. Excella. https://www.excella.com/insights/ml-model-security-preventing-the-6-most-common-attacks. Accessed August 29, 2023.

[36] AI Security White Paper(Oct 01, 2018). (n.d.). Huawei. https://www.huawei.com/en/trust-center/resources/ai-security-white-paper. Accessed August 29, 2023.

[37] Asif, M. (2022b, January 6). Data Poisoning: When artificial intelligence and machine learning turn rouge. Medium. https://medium.com/analytics-vidhya/data-poisoning-when-artificial-intelligence-and-machine-learning-turn-rouge-d8038f423922. Accessed August 29, 2023

[39] Short, A., La Pay, T., & Gandhi, A. (2019). Defending against adversarial examples. https://doi.org/10.2172/1569514

[40] Takyar, A. (2023, August 4). AI model security. LeewayHertz - AI Development Company. https://www.leewayhertz.com/ai-model-security/. Accessed August 30, 2023.

[41] OWASP AI Security and Privacy Guide | OWASP Foundation. (n.d.). https://owasp.org/www-project-ai-security-and-privacy-guide/. Accessed September 03, 2023.

[42] M. Bagaa, T. Taleb, J. B. Bernabe and A. Skarmeta, "A Machine Learning Security Framework for Iot Systems," in *IEEE Access*, vol. 8, pp. 114066-114077, 2020, doi: 10.1109/ACCESS.2020.2996214.

[43] National Cyber Security Center. (2022, August). Principles for the security of machine learning. https://www.ncsc.gov.uk/files/Principles-for-the-security-of-machine-learning.pdf

[44] Software Solution Architecture. (2023). How do you learn and improve software security architecture skills and knowledge for AI systems? www.linkedin.com. https://www.linkedin.com/advice/0/how-do-you-learn-improve-software-security. Accessed August 30, 2023.

[45] Kearns M, Roth A. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford: Oxford University Press; 2019.

[46] Kolamanvitha. (2022, March 30). Design patterns for Machine learning - towards data science. Medium. https://towardsdatascience.com/design-patterns-for-machine-learning-410be845c0db. Accessed August 30, 2023.

[47] Washizaki, Hironori & Khomh, Foutse & Guéhéneuc, Yann-Gaël & Takeuchi, Hironori & Natori, Naotake & Doi, Takuo & Okuda, Satoshi. (2022). Software-Engineering Design Patterns for Machine Learning Applications. Computer.

[48] Scott, Andrew & Solórzano, José & Moyer, Jonathan & Hughes, Barry. (2022). The Future of Artificial Intelligence. International Journal of Artificial Intelligence and Machine Learning. 2. 1.

[49] Walugembe, Francis Noah. (2023). The Future of Artificial Intelligence and Its Impact on Society.

[50] IBM Center for The Business of Government. (n.d.). The Future of Artificial Intelligence. David a. Bray.

[51] DeVerter, J. (2023, August 9). From Curation To Creation: How Ethical AI Can Shape A Responsible Future. Forbes. https://www.forbes.com/sites/forbestechcouncil/2023/08/09/from-curation-to-creation-how-ethical-ai-can-shape-a-responsible-future/?sh=4415dc261ea6. Accessed August 30, 2023.

[52] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, 103525. https://doi.org/10.1016/j.cviu.2022.103525

[53] Great Learning Team. (2022, November 21). Deepfake: What is Deepfake in Artificial Intelligence. Great Learning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/all-you-need-to-know-about-deepfake-ai/

[54] Hussein, Burhan & Haleem, Chongomweru & Siddique, Muhammad. (2021). The Future of Artificial Intelligence and its Social, Economic and Ethical Consequences.

[55] J.R. Slaby, Robotic Automation Emerges as a Threat to Traditiomal Low-cost Outsourcing, HfS Res. Ltd. (2012) 1–18.

[56] R. Perrault, Y. Shoham, E. Brynjolfsson, J. Clark, J. Etchemendy, B. Grosz, T. Lyons, J. Manyika, S. Mishra, J.C. Niebles, Artificial Intelligence Index 2019 Annual Report, 2019. https://hai.stanford.edu/sites/g/files/sbiybj10986/f/ai_index_2019_rep ort.pdf.

[57] Manyika, J. (2022, November 17). What do we do about the biases in AI? Harvard Business Review. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai. Accessed August 19, 2023.

[58] How do machines learn? A beginners guide. (n.d.). https://levity.ai/blog/how-do-machines-learn. Accessed September 01, 2023.

[59] Reddy, Sandeep. (2018). Use of Artificial Intelligence in Healthcare Delivery. 10.5772/intechopen.74714.

[60] G. Marcus, "Deep learning: A critical appraisal," arXiv preprint arXiv:1801.00631, 2018.

[61] López de Mántaras, R. (2019). Towards artificial intelligence: Advances, challenges, and risks. Metode Science Studies Journal , 9 , 119-125. doi:http://dx.doi.org/10.7203/metode.9.11145

[62] AI-Self Security. (n.d.). Huawei. https://www.huawei.com/kr/trust-center/ai-section. Accessed September 02, 2023.

[63] Hughes, A. (2023, January 13). Advancing human-centered AI: Updates on responsible AI research - Microsoft Research. Microsoft Research. https://www.microsoft.com/en-us/research/blog/advancing-human-centered-ai-updates-on-responsible-ai-research/. Accessed Septmber 02, 2023.

[64] Bell, C. (n.d.-b). How AI will impact the future of security. www.linkedin.com. https://www.linkedin.com/pulse/how-ai-impact-future-security-charlie-bell/. Accessed September 16, 2023.

[65] Smith, B. (2023). Meeting the AI moment: advancing the future through responsible AI. Microsoft on the Issues. https://blogs.microsoft.com/on-the-issues/2023/02/02/responsible-ai-chatgpt-artificial-intelligence/. Accessed September 04, 2023.

[66] Smith, B. (2023b, June 29). Advancing AI governance in Europe and internationally - EU Policy Blog. EU Policy Blog. https://blogs.microsoft.com/eupolicy/2023/06/29/advancing-ai-governance-europe-brad-smith/. Accessed September 04, 2023.

[67] OWASP AI Security and Privacy Guide | OWASP Foundation. (n.d.-b). https://owasp.org/www-project-ai-security-and-privacy-guide/#:~:text=They%20can%20be%20mitigated%20by,account%20when%20training%20a%20model. Accessed September 16, 2023.

[68] Reinforcement Learning Tutorial - JavatPoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/reinforcement-learning. Accessed September 14, 2023.

[69] Brendel, Alfred & Mirbabaie, Milad & Lembcke, Tim-Benjamin & Hofeditz, Lennart. (2021). Ethical Management of Artificial Intelligence. Sustainability. 13. 10.3390/su13041974.

[70] Buolamwini, J. &amp; Gebru, T.. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st

Conference on Fairness, Accountability and Transparency, in Proceedings of Machine Learning Research 81:77-91 https://proceedings.mlr.press/v81/buolamwini18a.html.

[71] Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review.

[72] Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. Advances in Neural Information Processing Systems.

[73] Lipton, Z. C. (2016). The Mythos of Model Interpretability. arXiv preprint arXiv:1606.03490.

[74] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[75] Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. Fordham Law Review.

[76] Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors: The Journal of the Human Factors and Ergonomics Society.

[77] Worldline en-global | Ever heard of the AI black box problem? (n.d.). https://worldline.com/en/home/main-navigation/resources/resources-hub/blogs/2021/ever-heard-of-the-ai-black-box-problem.html. Accessed September 15, 2023.

[78] Federspiel F, Mitchell R, Asokan A, et alThreats by artificial intelligence to human health and human existenceBMJ Global Health 2023;8:e010435.

[79] Lorenz-Spreen P , Oswald L , Lewandowsky S , et al . A systematic review of worldwide causal and correlational evidence on digital media and democracy. Nat Hum Behav 2023;7:74–101. doi:10.1038/s41562-022-01460-1

[80] Agudo U , Matute H . The influence of algorithms on political and dating decisions. PLoS One 2021;16:e0249454. doi:10.1371/journal.pone.0249454

[81] Javorsky E , Tegmark M , Helfand I . Lethal autonomous weapons. BMJ 2019;364:l1171. doi:10.1136/bmj.l1171

[82] Russel S . The new weapons of mass destruction? 2018. Available: https://www.the-security-times.com/wp-content/uploads/2018/02/ST_Feb2018_Doppel-2.pdf. Accessed August 20, 2023.

[83] AdminEticas. (2021).Nikon’s Face detection bias. Eticas Foundation. https://eticasfoundation.org/nikons-face-detection-bias/. Accessed August 30, 2023.

[84] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. *ArXiv*. /abs/1104.3913

[85] How artificial intelligence could widen the gap between rich and poor nations. (2020, December 2). IMF. https://www.imf.org/en/Blogs/Articles/2020/12/02/blog-how-artificial-intelligence-could-widen-the-gap-between-rich-and-poor-nations. Accessed August 30, 2023

[86] Barreno, M., Nelson, B., Joseph, A.D. et al. The security of machine learning. Mach Learn 81, 121–148 (2010). https://doi.org/10.1007/s10994-010-5188-5

[87] Ferrara, Emilio. (2023). Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies. 10.2196/preprints.48399.

[88] Karve, Swagat & Arpityadav, & Dutta, Prateek. (2022). Artificial Intelligence in Cyber Security. REST Journal on Emerging trends in Modelling and Manufacturing. 8. 10.46632/jemm/8/2/6.

[89] Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277.

[90] Fredrikson, M., Jha, S., & Ristenpart, T. (2014). Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 1322-1333.

[91] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2017 pp. 3-18. doi: 10.1109/SP.2017.41

[92] Segarra, L. M. (2021, June 9). Elon Musk: AI poses ‘Vastly more risk than North Korea.’ Fortune. https://fortune.com/2017/08/12/elon-musk-ai-poses-vastly-more-risk-than-north-korea/. Accessed August 30, 2023.

[93] Koolen, C., & van Cranenburgh, A. (2017). These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (pp. 12-22). Association for Computational Linguistics. https://aclanthology.org/W17-1602

[94] Timmermans, J., Stahl, B., Ikonen, V., & Bozdag, E. (2010). The Ethics of Cloud Computing: A Conceptual Review. In Proceedings: 2nd IEEE International Conference on Cloud Computing Technology and Science, CloudCom 2010 (pp. 614-620). IEEE Institute of Electrical and Electronic Engineers. https://doi.org/10.1109/CloudCom.2010.59

[95] Hernández-Orallo, J. (2017). The measure of all minds: Evaluating natural and artificial intelligence. Cambridge University Press. https://doi.org/10.1017/9781316594179

List of Figures

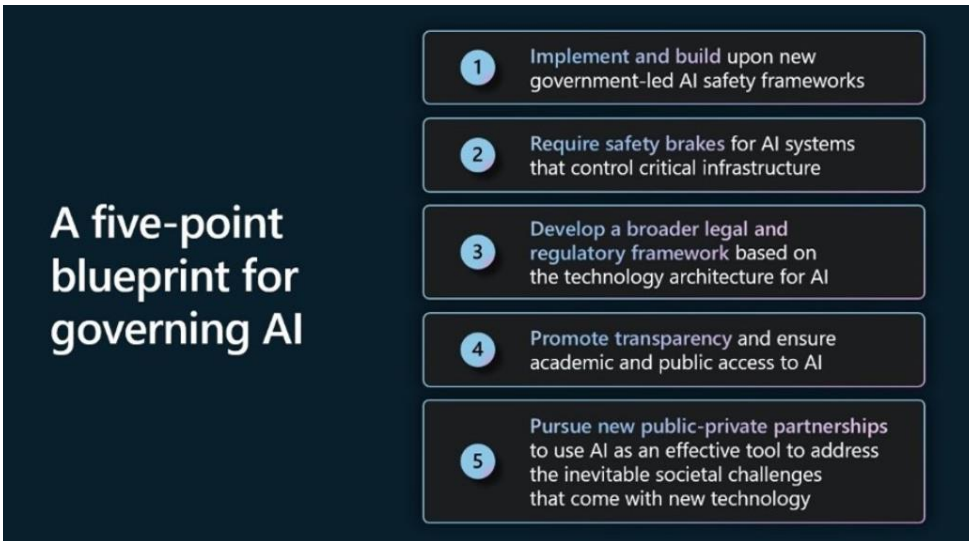
List of Tables

Appendix

5.1 Appendix A Other Possible Mitigations for a Secure AI

5.1.1 Microsoft AI Governance Blueprints

Microsoft published a whitepaper titled "Governing AI: A Blueprint for the Future" with the objective of examining the optimal approach to governing artificial intelligence (AI).



A five-point blueprint for governing AI

1. Implement and build upon new government-led AI safety frameworks
2. Require safety brakes for AI systems that control critical infrastructure
3. Develop a broader legal and regulatory framework based on the technology architecture for AI
4. Promote transparency and ensure academic and public access to AI
5. Pursue new public-private partnerships to use AI as an effective tool to address the inevitable societal challenges that come with new technology

Figure 13 Microsoft Blueprint for Governing AI [66]

The whitepaper suggests that A key element to ensuring the safer use of this technology is a risk-based approach, with defined processes around risk identification and mitigation as well as testing systems before deployment. The AI Act sets out such a framework and this will be an important benchmark for the future. the government would define the class of high-risk AI systems that control critical infrastructure and warrant such safety measures as part of a comprehensive approach to system management. New laws would require operators of these systems to build safety brakes into high-risk AI systems by design. The government would then oblige operators to test high-risk systems regularly. And these systems would be deployed only in licensed AI datacenters that would provide a second layer of protection and ensure security.

The AI Act acknowledges the challenges to regulating complex architecture through its risk-based approach for establishing requirements for high-risk systems. At the application layer, this means applying and enforcing existing regulations while being responsible for any new AI-specific deployment or use considerations. It’s also critical to advance the transparency of AI systems and broaden access to AI resources. While there are some inherent tensions between transparency and the need for security, there exist many opportunities to make AI systems more transparent. The AI Act will require that AI providers make it clear to users that they are interacting with an AI system. Similarly, whenever an AI system is used to create artificially generated content, this should be easy to identify. Important work is needed now to use AI to protect democracy and fundamental rights, provide broad access to the AI skills that will promote inclusive growth, and use the power of AI to advance the planet’s sustainability needs. [66]

5.1.2 OWASP Security Guidelines Against Adversarial Attacks [67]

To mitigate the capabilities of artificial intelligence (AI), one approach is to subject the model's behavior to either human control or automated oversight, wherein an additional algorithm establishes boundaries and constraints.

One potential approach to imposing constraints on the capabilities of artificial intelligence (AI) systems involves the reduction of privileges. This can be achieved, for instance, by refraining from establishing a connection between an AI model and an email service, so mitigating the risk of disseminating inaccurate or misleading information to external recipients.

In the context of data poisoning attacks, the manipulation of training data or the alteration of data labels can effectively influence the behavior of the model. The potential consequences of this action include compromising the integrity of the model or influencing its decision-making process in favor of the individual initiating the attack. The aforementioned assault exhibits characteristics akin to a Trojan horse, whereby the model's functionality appears normal under typical circumstances, but upon encountering particular modified inputs, it is compelled to make predetermined decisions. Language models such as ChatGPT, specifically LLMs (Large Language Models), generate source code by leveraging a vast training dataset comprising code samples sourced from many online platforms. However, it is important to acknowledge that this training dataset may potentially contain instances of security vulnerabilities or other forms of malicious behavior that have been introduced into the code. Countermeasures include the safeguarding of the data pipeline and the implementation of quality assurance measures for data.

Regarding evasion attacks, one notable technique involves deceiving models through the use of fraudulent input data. This attack can be executed through three distinct methods: 1) through the manipulation of the model input, commonly referred to as the black box approach, 2) by crafting input that is specifically meant to exploit the model parameters, known as the white box approach, and 3) by utilizing data poisoning as the foundation for generating the input. The most effective approach to mitigating adversarial attacks involves the utilization of robust-performing models, in conjunction with various mitigation strategies. These strategies encompass measures such as poisoning mitigation, restriction of access to model parameters, exclusion of confidence values from the output, throttling, as well as the implementation of monitoring and detection mechanisms to identify manipulation types, including physical patches in images. Furthermore, it is possible to enhance the training process by using adversarial samples, so increasing the model's resilience against manipulated input. This objective can also be accomplished by employing a method known as randomized smoothing.

In the context of membership inference attacks, the objective is to ascertain whether a given data record, such as an individual, was part of the training dataset used to train a model, using just black-box access to the model. The issue at hand can be categorized as a non-repudiation challenge, when an individual is unable to refute their affiliation with an organization of a sensitive nature. Overfitting, which refers to the phenomenon where a model becomes excessively specialized in recognizing patterns within the original training set, poses an increasingly significant concern. One potential approach to mitigating overfitting involves employing strategies such as constraining the model's complexity, increasing the size of the training dataset, or introducing noise into the training data.

Furthermore, model invasion assaults refer to attacks that involve interacting with or evaluating a model. Through these attacks, it becomes possible to estimate the training data with varied degrees of accuracy. This issue becomes more problematic when the training data includes confidential or copyrighted information. In order to adhere to best standards, it is advisable to refrain from including sensitive or personal data within the training set. Additionally, it is crucial to prevent models from overtraining, which can be achieved by employing adequately large training sets. Implementing restrictions on access to the model can serve as an additional measure to deter unauthorized manipulation or examination. The field of generative artificial intelligence also has its own set of issues in this context. Query-answer models have a potential risk of generating responses that are derived from sensitive training data, potentially leading to memorizing. On the other hand, Generative AI systems have the capability to generate text, images, or videos that may contain sensitive or copyrighted content. One noteworthy scenario arises when a Generative chat system is intentionally exploited to disclose confidential sensitive information, as exemplified by the instance of Bing in February 2023.

[104]