NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

The Landscape of Artificial Intelligence Ethics:

Analysis of Developments, Challenges, and Comparison of Different Markets

Simon Natrup

Dissertation presented as a partial requirement for obtaining a Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

THE LANDSCAPE OF ARTIFICIAL INTELLIGENCE ETHICS: ANALYSIS OF DEVELOPMENTS, CHALLENGES, AND COMPARISON OF DIFFERENT MARKETS

by

Simon Natrup

Dissertation presented as a partial requirement for obtaining a Master's degree in Information Management, specialization in Information Systems and Technologies Management

Coordinator: Prof. Dr. Vitor Duarte dos Santos

ABSTRACT

Artificial Intelligence has become a disruptive force in the everyday lives of billions of people worldwide, and the impact it has will only increase in the future. Be it an algorithm that knows precisely what we want before we are consciously aware of it or a fully automized and weaponized drone that decides in a fraction of a second if it may strike a lethal attack or not. Those algorithms are here to stay. Even if the world could come together and ban, e.g., algorithm-based weaponized systems, there would still be many systems that unintentionally harm individuals and whole societies. Therefore, we must think of AI with Ethical considerations to mitigate the harm and bias of human design, especially with the data on which the machine consciousness is created. Although it may just be an algorithm for a simple automated task, like visual classification, the outcome can have discriminatory results with long-term consequences. This thesis explores the developments and challenges of Artificial Intelligence Ethics in different markets based on specific factors, aims to answer scientific questions, and seeks to raise new ones for future research. Furthermore, measurements and approaches for mitigating risks that lead to such harmful algorithmic decisions and identifying global differences in this field are the main objectives of this research.

KEYWORDS

Machine Learning Ethics; Artificial Intelligence Ethics; Machine Learning Bias; Machine Learning Discrimination; Trustworthy AI; Artificial Intelligence Principles; Artificial Intelligence Guidelines

TABLE OF CONTENTS

1. INTRODUCTION

1.1. CONTEXT

Artificial Intelligence (AI) is increasingly **disrupting** various areas of everyday life and industries at high velocity, whether in autonomous driving, medical diagnostics, or virtual assistants. (Cf. Litjens et al., 2017, p. 60; Marina et al., 2018, p. 559; Lopatovska et al., 2018, p. 2 f.) Additionally, algorithms and learning systems play an increasingly important role in many other parts of medicine, in road traffic, in decisions about the allocation of jobs or loans, renting of apartments, or even the choice of a partner, and not least in the context of warfare (cf. Rahmani, 2011; Ronneberger et al., 2015; Esteva, 2017, p. 115; Katzenmeier, 2019, p. 259). Although most of the use cases just mentioned are generally positive, visions of the future with AI often have dystopian characteristics. In the worst scenarios, machines take over the world, dominate us, enslave us, or wipe us out. (Cf. Broman, 2017; Joy, 2000)

The breakthrough of AI is due to the rapid increase in computing power combined with the availability of large amounts of inexpensive data (cf. Bartlett, 2018). On the algorithmic side, there are far-reaching developments in the field of Machine Learning (ML). ML can be described as a subarea of AI and simultaneously as a driving force for its success. (Cf. Awad & Khanna, 2015, p. 1; Kubat, 2017, p. ix). In addition to the increase in raw computing power as an enabler of AI, there is also the emergence of specific AI hardware, e.g., the neuromorphic chip Loihi by Intel (cf. Davies et al., 2018; Batra et al., 2018). This relatively new AI chip market will be worth 73 billion USD by 2025 (cf. Technavio, 2021). The global AI market is worth 327.5 billion USD in 2021 and is expected to grow to 554.3 billion USD in 2026 (cf. IDC, 2021). By 2030, the global AI market will grow to 15.7 trillion USD (cf. Rao et al., 2017).

A study conducted by Ipsos in 2020 for the European Commission concluded that 42% of enterprises already use at least one AI technology and that 18% have plans to adopt AI technologies in the next two years (cf. IPSOS, 2020). As a result of the advances and the rising adoption of AI in almost all areas of society, the technology is also becoming a field of action for policymakers. This includes promoting and regulating new technologies and addressing vague fears in the population about potential adverse effects on people and society. Moreover, the policy itself may also change if AI becomes the basis of political decision-making processes. (Cf. Rieder & Simon, 2016)

Policymakers are especially gaining interest since the application of AI systems becomes more intertwined with politics and governments. The opportunities for the utilization of AI in that sector are manifold and range from advisory and recommendation systems to complex decision-making systems. Especially in the latter case, i.e., when decisions about the distribution of goods and services and hazards and risks are made with recourse to AI-based systems, the question of their potential for **unfairness** and **discrimination** inevitably arises. (Cf. Binns, 2018a; Binns et al., 2018b)

It is a common phenomenon that new ethical considerations and challenges follow the spread of disruptive technologies. That has been seen before with, i.e., nuclear power (cf. Taebi et al., 2012), genetically modified organisms (cf. Hielscher et al., 2016), or biotechnology and bioengineering (cf. Munshi & Sharma, 2018).

In the case of AI, one of the ethical problems that have drawn attention is that AI systems can amplify all kinds of **racial** and **gender biases** (cf. Caliskan et al., 2017). Besides amplified biases that often come unknown and unwanted, there is also the **ethically questionable use** of AI. Google, a company that once proclaimed "Don't be evil" as one of its main principles, faced a backlash by its employees for taking part in an initiative that explored the use of AI for weapon systems, i.e., the Maven project. (Cf. Google, 2004; Maas, 2019) The idealized principle was changed to "do the right thing" in 2014 (cf. Google 2014). The backlash was successful, Google did not renew the contract with the US Department of Defense in 2018, but the interest in AI's ethical or unethical use accelerated further (cf. Google 2014; Maas 2019).

Besides gathering attention from AI tech companies, academia, and civil rights groups, AI Ethics has also drawn attention even from the theological sphere. The Vatican published a statement regarding the ethical commitment of AI in 2020 during a conference called the "renAIssance Call for an AI Ethics." (Cf. Rome Call, 2020) Participants at this and similar discussions try to define ethical AI principles, explore why AI biases occur, and try to identify solutions to prevent algorithms from systematically favoring or disadvantaging certain citizens based on their gender, origin, or religion.

1.2. STUDY OBJECTIVE

The main objective of this thesis is to analyze the landscape of AI ethics. This includes an introduction to the fundamentals of AI, ethics, and AI ethics. Furthermore, it is the objective to explore development and challenges, including biases, discriminations, and harms resulting from it. Based on the findings, different markets will be analyzed and compared to specific factors defined based on a systematic literature review.

The goals are defined as:

Goal 1 - Identifying and analyzing challenges in AI Ethics.

Goal 2 - Identifying harm and discrimination through the lack of AI Ethics.

Goal 3 - Identifying and comparing AI Ethics in different markets.

Goal 4 - Raising questions for further research.

Research questions are defined as:

Question 1: What are the developments in the field of AI Ethics?

Question 2: What are the challenges in the field of AI Ethics?

Question 3: How do different markets adjust to AI Ethics?

Question 4: What is the view of industry leaders on AI Ethics?

1.3. STUDY RELEVANCE AND IMPORTANCE

Various industries incorporate AI applications in multiple sectors, e.g., supply-chain management, manufacturing, marketing, service development, and risk assessment (cf. Wiggers, 2019).

The application of AI opens a multitude of new **opportunities** for government and society. It promises significant gains in effectiveness and efficiency in the execution of government tasks in the areas of education, mobility, health, etc. (cf. Eggers et al. 2017, p. 2 ff.). Due to the potential for opportunities and **transformation** for governmental use and for accelerating economic power, AI has moved far up the political agenda of many countries (cf. Dutton, 2018).

Since AI as a cross-cutting technology is characterized by an almost universal range of applications and is still in its infancy, both in the development process and in use, it is essential to critically examine the conflicts that are already occurring today and to accompany the development and dissemination of AI technologies accordingly (cf. Weyerer & Langer 2019, p. 509 f.).

Society needs to find answers in the field of AI Ethics – hereafter referred to as AIE. There needs to be a consensus about which areas of life decisions can be transferred to an algorithm, which correlations and categorizations are acceptable and discriminatory.

This research seeks to help organizations better understand the **opportunities** and **challenges** that emerge through AI's disruptive impact regarding Ethics. It will enable regulators to find feasible **policies** that **mitigate risks** but do not impede the potential of the technology. It is essential to find regulations for different fields of application to assure that not only one interest group benefits. The balance between individual data privacy and opportunities for the collective based on large amounts of data combined with AI must be established.

The analysis of how the technology is already used in organizations will support companies by identifying ethical gaps that need to be mitigated. This will be done by showing the main problems to help future projects and companies. By identifying differences in how markets approach ethical issues and challenges and which measures work out, it will be possible to give **valuable recommendations**. Thus, helping the **industry** and **governments** adapt and benefit during the new age of AI while balancing ethical considerations within the **society**.

Data regarding scientific publications with the exact keywords used for this research has been analyzed to demonstrate the growing interest in the field and therefore show study relevance and importance. It can be shown through the data that the interest in the subject is increasing tremendously over the last five years. Created with Clarivate's Web of Science. (Clarivate, 2021)

Keyword: Artificial Intelligence Ethics

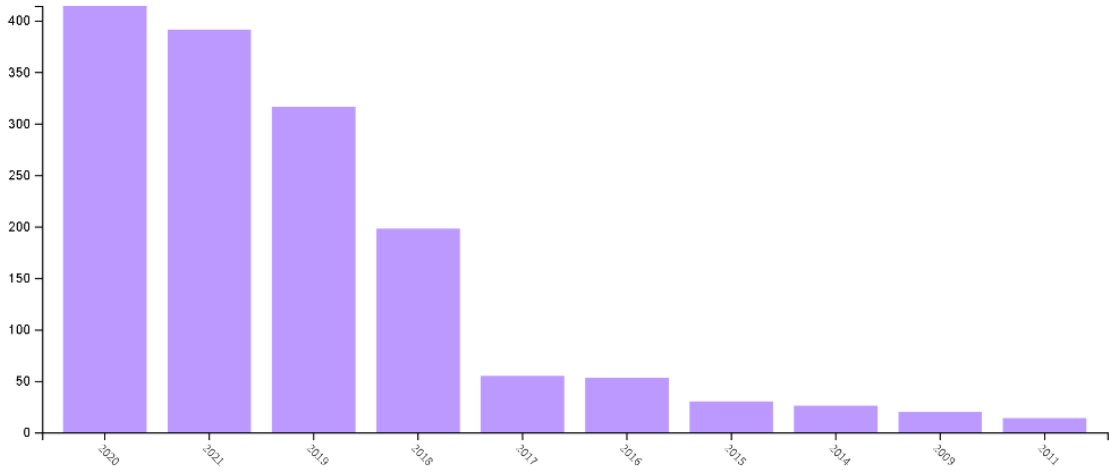

Figure 1 - Artificial Intelligence Ethics (Carivate, 2021)

Keyword: Trustworthy AI

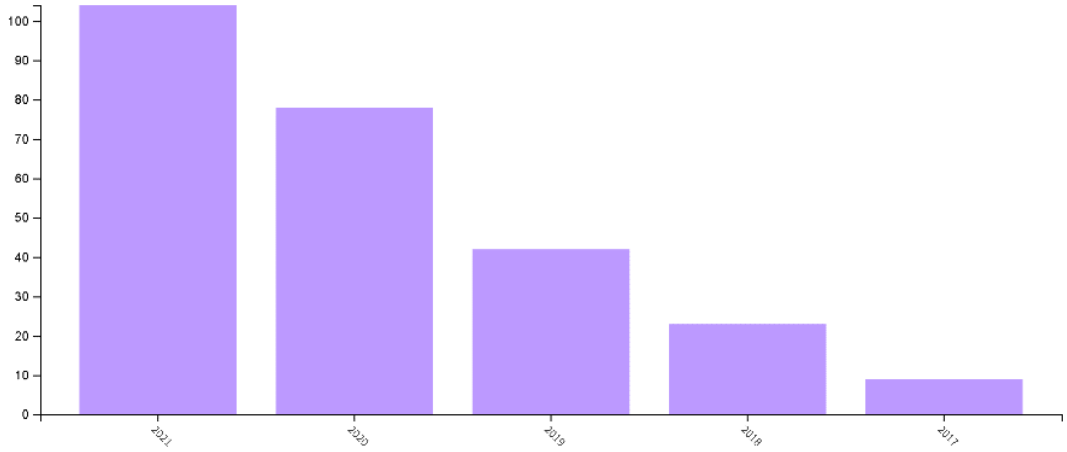

Figure 2 - Trustworthy AI (Carivate, 2021)

This research will bring **scientific value** by identifying current research gaps and accumulating conclusions from various papers. The author aims to publish it in a scientific journal to reach a bigger audience and to have an impact on the academic discussion.

1.4. METHODOLOGICAL OUTLINE

The methodology for this research was composed of four phases: Exploration Phase, Analytical Phase, Execution Phase, and Conclusive Phase.

The outcome of the Exploration Phase was a detailed assessment of the literature and scientific developments. The Analytical Phase accomplished the intermediate goals (1) to (4) by facilitating a state-of-the-art literature review, analysis of results and challenges, analysis and comparison of different markets, which are the EU, US, and China.

Finally, the Conclusive Phase aims to fully answer the scientific questions (1) to (4). Furthermore, the last phase discusses and concludes the results of all the other phases and facilitates an outlook into future developments. The following graph illustrates the procedure to reach the intermediate goals that have been defined in Figure 3:

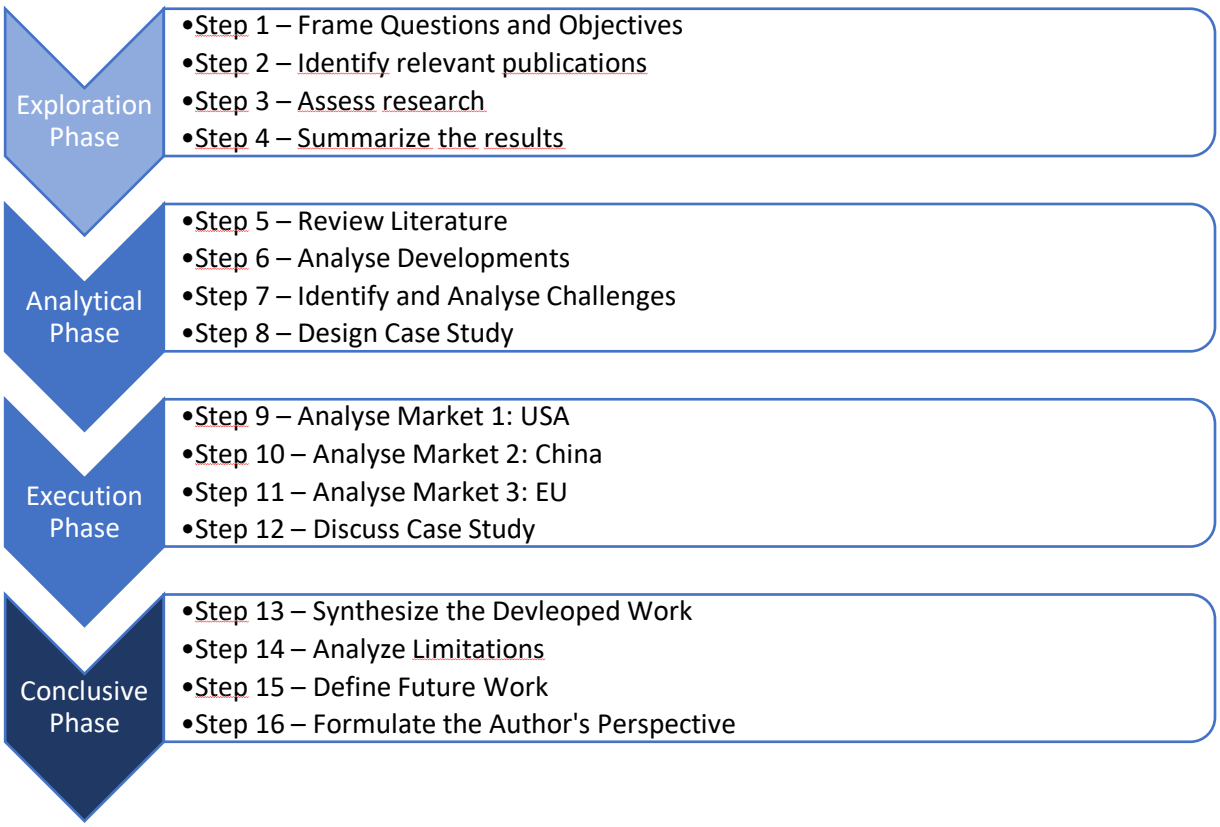

Figure 3 - Methodology

To reach the objective, it is essential to identify important leaders in the field of AI from the three markets in academia and the industry. Their perspectives regarding the subject will almost certainly have an impact on further developments. Therefore, their views might give us insights into the future. The literature review is focused on the overall subject of everything surrounding Artificial Intelligence Ethics, while the Case Studies focus on specifics about the different markets. Those specifics will be defined as factors as an outcome of the literature review.

2. LITERATURE REVIEW

2.1. LITERATURE REVIEW METHODOLOGY

This thesis follows a literature review approach that analyzes state-of-the-art scientific research to develop a structured overview with detailed insights. With that aim, it mainly considers the most recent scientific publications. Older literature was also being used when it was necessary and to describe the basics. At the beginning of this work around 50 scientific articles have been selected, which were filtered concerning various characteristics.

Especially publications from journals with a high impact factor have been considered. The relevance of the papers was, amongst other things, defined by how often it has been cited. Furthermore, distinguished experts in this field were identified and quoted when it was helpful for the objective of this thesis. Journals cited in this thesis are, e.g., Oxford Journal of Legal Studies, Journal of Defense Management, Journal of Ethics, International Journal of Communication, Journal of Database Management, International Journal of Public Administration, Journal of Artificial Intelligence Research, Nature, and Nature Machine Intelligence.

The research also relies on articles and interviews by popular media outlets, e.g., CNBC, Engadget, Wired, and The New York Times.

2.1. ARTIFICIAL INTELLIGENCE

2.1.1. AI Fundamentals

Artificial intelligence is the name of a subfield of computer science that has existed since the mid-1950s. It focuses on the automation of complex problem-solving processes with the aid of computer programs. The term artificial is used in contrast to natural, i.e., human intelligence. The definition of human intelligence varies greatly depending on the scientific discipline (Gardner, 1999; Piaget, 2000; Maltby et al., 2011), which is why some researchers prefer to speak of intelligence in terms of information processing (Trappl, 1986), extended intelligence (Ito, 2017), or designed intelligence (Davies, 2017). However, the term artificial intelligence is more common.

Specific AI experiences are understood as if they correspond to cognitive, emotional, or other competencies and capabilities and are based on the distinction between **strong and weak AI**. According to **John McCarthy**, one of the most influential scientists in AI, its goal is to build machines that behave intelligently. The focus for him was on weak AI, which does not strive to create consciousness. For him, the claim for weak AI is to focus on concrete applications that require intelligent solutions. (Cf. McCarthy, 2007, p. 2-13) Weak AI is merely concerned with simulating specific competencies in artificial systems (cf. Searle, 1980). In other words, it is concerned solely with the functions of human intelligence being mapped by machines, e.g., using neural networks. The goal here is to develop various methods for problems such as Natural Language Processing (NLP) (cf. Chowdhury, 2003), software development for automated vehicles (cf. Cox & Wilfong, 2012), analysis systems based on large data sets, and the development of intelligent navigation systems (cf. Herbert, et al. 2012).

Strong AI means machines equipped with intelligence, consciousness, and autonomy in the genuine sense of the word. Representatives of this approach strive for the complete machine emulation of

mental processes such as thinking, learning, or problem-solving. (Cf. Nilsson, 2010) Stuart Russell and Peter Norvig (2003, p. 947), throughout their definition of weak and strong AI, noted that "the assertion that machines could act intelligently (or, perhaps better, act as if they were intelligent) is called the weak AI hypothesis by philosophers, and the assertion that machines that do so are thinking (as opposed to simulating thinking) is called the strong AI hypothesis."

Both variants' premise is that beliefs are a kind of information, the reasoning is a kind of computation, and motivation is a kind of feedback and control (cf. Pinker, 2015, p. 31).

Russell and Norvig (2003, p. 947) furthermore argue that "[m]ost AI researchers take the weak AI hypothesis for granted, and do not care about the strong AI hypothesis – as long as their program works, they do not care whether you call it a simulation of intelligence or real intelligence."

In the literature, reference is made to Thomas Hobbes, among others, and his thesis formulated in Computatio, Sive, Logica. "I understand computation. And to compute is to collect the sum of many things added together at the same time, or to know the remainder when one thing has been taken from another. To reason, therefore, is the same as to add or to subtract." (Hobbes, 1655, p. 3)

Which could be rephrased and summarized as 'thinking is calculating.' This quotation is used as a basis to support the thesis that the human brain or cognitive abilities such as thinking can be artificially reproduced. Thus, the computer scientist Herbert A. Simon, even before the founding of the field of AI, described: "Any rational decision may be viewed as a conclusion reached from certain premises. (…) The behavior of a rational person can be controlled, therefore, if the value and factual premises upon which he bases his decisions are specified for him." (Simon, 1944, p. 19)

The assumption that essential functions of human cognitive performance can be implemented by machines, both in terms of hardware and software, formed the basis of all research projects during AI development (cf. Feigenbaum & Feldman, 1963; Feldmann, 2003).

In this sense of the term, the development of AI can be equated with the imitation of human cognitive characteristics utilizing computer technology. However, the feasibility of this endeavor - and thus the danger that **machines could replace humans** - is controversial. More realistic concern about developments in AI that already have an impact on people's freedom, for example, when personal data is collected and analyzed for economic purposes. (Cf. Kurzweil, 2005)

Michael I. Jordan, who was ranked as the most influential computer scientist in 2016 by Semantic Scholar, argues that "people are getting confused about the meaning of AI in discussions of technology trends" when they think "that there is some kind of intelligent thought in computers that is responsible for the progress, and which is competing with humans." In his observation, "We don't have that [kind of technology yet], but people are talking as if we do." (Jordan, 2021)

2.1.2. Machine Learning

AI is mainly driven by **Machine Learning (ML)**. ML is a subfield of the interdisciplinary field of AI, and its main features can be traced back to the work of **Alan Touring** during the 1950s (cf. Turing, 1950, pp. 433-460). **Arthur Samuel**, one of the pioneers in AI, defined ML as an area of research that enables computers to learn without being explicitly programmed to do so (cf. Samuel, 1959, p. 211).

ML aims to learn in an automated way, based on observations of the natural world, without explicit logic or rules. From the observation based on the processing and classification of training data, the experience should emerge. The accumulated experience becomes the basis for an automated improvement in accuracy, thus leading to the actual learning effect. (Cf. Khan et al., 2018, p. 5)

ML approaches can be categorized differently according to their flow behavior. Those behaviors can be differentiated broadly into supervised and unsupervised learning methods. **Supervised learning** involves processes in which associations are extracted between given known attributes and unknown attributes. Training data is used for the learning effect, which consists of input and output values. This results in a model that can generate associations with previously unknown input values based on experience. In such models, the performance depends on the variance and amount of training data. High variance and amount of training data leads to better generalization and increased ability to generate associations with unlabeled input values. (Cf. Awad et al., 2015, p. 4) Decision trees, random decision forests (RDFs), artificial neural networks (ANNs), and support vector machines (SVMs) are among the supervised learning methods (cf. Khan et al., 2018, p. 5).

Unsupervised learning includes approaches that group instances without a predefined attribute. Accordingly, such a model acts autonomously in classifying input values based on distinguishing features. (Cf. Awad et al., 2015, p. 4). These include the hidden Markov model (HMM) and the Gaussian mixture model (GMM) (cf. Khan et al., 2018, p. 5). There are also approaches, such as the Restricted Boltzmann Machine (RBM), which can be trained, supervised or unsupervised depending on the task, so a clear distinction is not possible (cf. Larochelle & Bengio, 2008, p. 536).

Between supervised and unsupervised learning, there is also **partially supervised learning**. In this approach, a large amount of training data is generally available, and only part of it is labeled. The remaining labels are then learned unsupervised. (Cf. Khan, 2018, p. 6)

2.1.3. AI Input, Design, and Database

The basis of all AI decisions is the human programming of the software and the given database, which is also determined by human influences, e.g., in the context of data collection and processing. In analogy to infants' learning process, which are born free of prejudice, it is the parents, the environment, and corresponding experiences that shape prejudice and discrimination. In the words of AI expert Kate Crawford: "Like all technologies before it, artificial intelligence will reflect the values of its creators." (Cf. Crawford, 2016, p. 11)

Human's design artificial learning and processing patterns with corresponding emotions, values, fears, knowledge gaps, and biases. In doing so, they can be consciously shaped as political tools (e.g., political Twitterbots) or unconsciously shaped by developers' preferences (e.g., AI-assisted personnel selection). The results are applications that discriminate against corresponding groups of people in a sexist or racist way. AI solutions thus often do not fulfill the objectivity expectations placed on them. In addition to the technical design of the AI, which results from programming and design by humans, it is the database (such as training data sets) that forms the basis for all evaluations and further processing of an AI. The learning and performance of AI depend crucially on the quantity and quality of the data available and accessible. However, the data never reflects the whole reality. Hence, a common reason for wrong conclusions of an AI application is usually an incomplete data set. In

addition, some data sets contain hurtful statements, for example, in the form of racist reports or discriminatory speech. (Cf. Weyerer & Langer, 2019, pp. 509 ff.)

Most data that medical ML models in the US are trained on comes from just three states, which are Massachusetts, New York, and California. There is little to no representation from the other 47 states. These three states may have economic, social, and cultural features that are not representative of the entire country. Therefore, ML models that are primarily trained on data from these states may generalize inadequately, which is an established risk when implementing diagnostic algorithms in new places. (Cf. Kaushal et al., 2020)

2.1.4. AI Learning and Processing

While algorithms so far have mainly made decisions based on comprehensible programming codes and therefore enabled traceable discrimination, the increased use of **deep learning** changes that. It is not always possible to trace the basis of which the algorithmic decision is made when it is done by a deep neural network. Hence, decision-making processes of such algorithms are often described as a **black box**. (Cf. Castelvecchi, 2016, pp. 20-23.)

The learning process of modern AI technologies is based on the evaluation of large data sets (**Big Data**) and the identification of meaningful structures. In addition, AI technologies often take their cue from human decisions and try to reproduce them based on historical data and, if necessary, replicate them in new situations (cf. Arel et al., 2010, pp. 13 ff.). The process of AI learning and decision-making is divided into three steps. First, the data basis is recorded, recognized, and adopted (recognition). Then, the data sets are analyzed and checked for patterns (understanding). Finally, the data is exploited to perform an appropriate task. In the process, the recognized patterns and dependencies are checked for the task and abstracted accordingly (producing). In all three of those phases of AI learning and decision-making, there can be causes for discriminative results. (Cf. Weyerer & Langer, 2019)

As described previously, input is a critical factor in AI application outcomes. If the data input is biased or insufficient, analysis results will also be incomplete or incorrect, which may lead to individuals or groups being unfairly discriminated against. The database made available for ML is thus the first starting point for avoiding discriminatory AI results. The understanding phase is determined by the programming of the AI application. Here, too, corresponding biases may be present, which must be humanely critically examined and adjusted to prevent potentially discriminatory results. Finally, the result of AI is not discriminatory until it is produced or published accordingly by the system. (Cf. Weyerer & Langer, 2020)

2.1.5. Race for AI Supremacy

A **race for AI supremacy** has long established itself with a general In- and outgroup thinking. Like in many other areas there is a deep competitive thinking between the US, China, and the EU – more even between the West (US, EU) and China – which comes with advantages and disadvantage. Competitors are easily seen as enemies or at least threats when it comes to a complex and disruptive technologies like AI. (Cf. Cave & ÓhÉigeartaigh, 2018)

Hagendorff argues that the **AI race** can be reframed into a global cooperation for safe and beneficial AI by abandoning such in- and outgroup thinking (cf. Hagendorff, 2020). The current AI race stands in contrast to the idea to develop a so-called **AI4people** or **AI for Global Good** (cf. Floridi et al., 2018).

According to Nicolas Chaillan, the former DoD software chief, the US has already lost the AI race to China. He declared in October 2021 that China is heading for global dominance (cf. Chaillan, 2021).

The previous and the current section already slightly emphasized on bias and discrimination, which are ethical risks of AI, to underline the relevance of their subjects. In the following section the fundamentals of ethical theory will be laid out to fully comprehend the nuances of AIE.

2.1. ETHICAL THEORY

The Merriam Webster Dictionary defines ethics as "the discipline dealing with what is good and bad and with moral duty and obligation", "a set of moral principles", "the principles of conduct governing an individual or a group ", "a guiding philosophy", and "a consciousness of moral importance" (Merriam-Webster, 2021). Ethics can also be defined as the philosophical study of morality. The word is often used synonymously with morality, and sometimes it is used as the moral principles of a particular individual, a group, or a tradition. (Cf. Audi, 1999, p. 883) For example, Immanuel Kant's ethics or Christian ethics (cf. Geisler, 1989; Sullivan, 1994).

Ethics, along with epistemology, metaphysics, and logic, is one of the main fields of philosophy. It can be divided into the metaphysics of moral responsibility, the general study of right action, the general study of goodness, metaethics, applied ethics, and moral psychology. Many studies in ethics, particularly those that validate or construct whole systems of ethics, are interdivisional - occurring between or involving two or more divisions. Those divisions facilitate the identification of different schools, movements, and problems within the discipline. The main business of ethics is constituted by the general study of right action and the general study of goodness. This results in the substantive questions of what objectives we should aim for and what moral principles should guide our decisions and aspirations. (Cf. Audi, 1999, p. 883 f.)

Normative ethics elaborates and examines universally valid norms and values as well as their justification. It is the core of general ethics. As a reflective theory of morality, it evaluates and judges what is good and right. (Cf. Churchill, 1999; Paul & Elder, 2019; Resnik, 2011) Such as deontological theories, e.g., Kant, and consequentialist theories, e.g., utilitarianism.

Applied ethics builds on normative ethics. It expresses itself as individual and social ethics as well as in area ethics for specific areas of life, for example medical ethics or business ethics. Ethics committees, councils and institutes develop standards or recommendations for action in specific areas. (Cf. Moor, 2020; Anderson & Anderson, 2011, p. 1)

Ethics always stress the danger of an artificial differentiation between in- and outgroups (cf. Derrida, 1967). Outgroups are perceived de-individualized, subjected to devaluation, and can become victims of violence just for being seen as the other (cf. Mullen & Hu, 1989; Vaes et al., 2014).

Nearly 40 million people from 233 countries went through a conjoint analysis related to the self-driving car, in which they were asked to weigh nine moral preferences: for example, whether the car should be more likely to protect people or animals, more likely to protect women or men, more likely to protect older people or younger people, more likely to protect rich people or poor people, and so on. The ethical theory underlying this experiment corresponds to **utilitarianism**. Utilitarianism was invented by Jeremy Bentham and John Stuart Mill in the 18th century (cf. Mill, 1863; Bentham, 1879; Lu, 2020). It postulates that an action is judged to be morally right if its consequences lead to the

greater good. In other words, that ethical problems can be solved according to the rule of making decisions in terms of the greatest possible happiness for the greatest possible number of people. And this happiness is determined in utilitarianism by weighing the advantages from a decision against the corresponding disadvantages. According to this, "ethical" algorithms would have to decide in such a way that the sum of the benefits outweighs the disadvantages. To know how important advantages and disadvantages are, e.g., in the scenario with the self-driving car, one could quantify social expectations, and if necessary, also with different weights for different cultures. At least this is what the scientists around the Moral Machine Experiment ("quantifying societal expectations about the ethical principles that should guide machine behaviour") recommend. (Cf. Anderson, Anderson, & Armen, 2005; Awad et al., 2018)

Deontological ethics, also known as duty-based ethics, on the contrary argues that actions should be evaluated not based on their expected outcomes, but on what people do. Duty-based ethics teaches that actions are wrong or right regardless of the bad or good consequences that could be produced. Under this form of ethical theory, one can't justify an action by showing that it produced good consequences. Kant's ethic is deontological. The first part of Kant's Categorical Imperative imposes on actors the duty to act only according to such a maxim that one would wish to become a general law. Maxims are rules and principles of personal action. (Cf. Kant, 1786) A utilitarian weighing of characteristics of equal people is hardly compatible with the culture of duty shaped by Kant (cf. Powers, 2006).

Another notable theory is the principle of the **Double Effect**. This theory states that if doing something morally good has a morally bad outcome, it is ethically acceptable to do it providing that the bad side-effect was not intentional. (Cf. Bonnemains et al., 2018)

The sociologist Ulrich Beck (1988, p. 194) noted that "[i]n the model of the objectified sciences, ethics plays the part of a bicycle brake on an intercontinental airplane".

2.1. Artificial Intelligence Ethics

2.1.1. Introductory Considerations

Depending on whether one uses utilitarianism, deontological ethics or the principle of double effect, a dilemma where an **AI powered drone** to take out a missile threatening an allied ammunition factory is unexpectedly alerted to a second threat - a missile heading towards some civilians, the decision outcome will be different. The drone must decide whether to continue its original mission or take out the new missile to save the civilians. (Cf. Bonnemains, et al. 2018) Centuries old ethical theories cannot agree on the merits of various approaches. Companies, governments, and researchers will find it even more difficult to decide which system to use for artificial agents (cf. Bogosion, 2017). Societies' or individual people's personal moral judgements can also differ widely when faced with moral dilemmas. The is especially that case when they are confronted with politicized issues such as economic inequality and racial fairness. (Cf. Greene et al., 2001) Bogosian (2017) argues that instead, we should design machines to be **fundamentally uncertain about morality**.

The perceptions and assessments of algorithms are assumed to be influenced by a plurality of factors and are not solely determined by the technical functionality. To put it differently: If an algorithm classifies without discrimination in the aggregate (factual fairness), this does not automatically mean that it is perceived as fair at the individual level (perceived fairness). However, people's perceptions are an important point of reference in the use of AI in politics and other areas regarding the legitimacy of decision-making procedures and decision outcomes. Factual and perceived fairness are both central to the stability of the democratic order in the digital society (cf. Verba 2006). It can be expected that the more the autonomy of AI increases, the more problematic and complex the ethical attribution of its behavior will be (cf. Neuhäuser, 2012, p. 24).

"**AI governance** moves from the realm of abstract principles into the world of **mass politics**." (Zhang & Dafoe, 2020) The lack of **transparency** risks undermining meaningful **control** and **accountability**, which is a problem when these systems are applied in the context of decision-making processes that can have significant human rights implications (cf. Koene et al. 2019). The question arises as to whom the decisions of the machine are to be attributed. It could be problematic to primarily attribute it to the human being who approves the decision. It is not uncommon for the ultimate decision-making human to be overwhelmed when the difficulty lies in deciding against an algorithmic suggestion that is known to be based on a very large amount of information and immense computing power. In this respect the very concept of **responsibility** has being fundamentally changed by the new developments at the moral and legal level. In any case, it can be problematic to focus primarily on the person approving the decision, the "human in the loop". (Cf. Sharkey 2016, pp. 23-38; Zanzotto, 2019, pp. 247 f.)

Human in the loop (HITL) can be divided into (1) Assisted Intelligence (helping people to perform tasks faster and better) and (2) Augmented Intelligence (helping people to make better decisions). **No human in the loop** (NHITL) can be divided into (1) Automated Intelligence (automation of cognitive/manual and routine/non-routine tasks) and (2) Autonomous Intelligence (automating decision making processes without human intervention). (Cf. Rao et al, 2017) There is also the so-called **society in the loop** (SITL). The term was forged by Rahwan (2017) and "combines the HITL control paradigm with mechanisms for negotiating the values of various stakeholders affected by AI systems, and monitoring compliance with the agreement." In short: Huaman in the loop + Social Contract.

In the political and scientific discourse, there are especially questions about how the development and application of AI as well as its consequences are to be evaluated ethically and morally, or how basic ethical values can be integrated into AI applications (cf. Lin et al., 2012, pp. 3 f.; Anderson & Anderson 2007, pp. 15 ff.; Wirtz et al., 2019, pp. 596 f.).

Siau & Wang (2021, p. 76) have created a concise and thought through distinction between **Ethics of AI and Ethical AI** (see Figure 4). They also illustrated in their research that the Ethical AI is a result of Ethics of AI (see Figure 5).

	AI	Human	Society
Ethics of AI	Principles of developing AI to interact with other AIs ethically	Principles of developing AI to interact with human ethically	Principles of developing AI to function ethically in society
Ethical AI	How AI should interact with other AIs ethically?	How AI should interact with humans ethically?	How AI should operate ethically in society?

Figure 4 - Ethics of AI and Ethical AI Distinction (Siau & Wang, 2021, p. 76)

Figure 5 - Ethics of AI to Ethical AI (Siau & Wang, 2021, p. 76)

Major **risks and challenges** exist in connection with the application of AI, which can be of technological, legal, social, and ethical nature (Wirtz et al. 2019, pp. 604 ff.).

Negative effects on society and the individual in addition to the concrete decision-making situation. Increasing unhappiness through interaction with social networks, which are algorithmically built to hook users on their platforms. (Cf. Ward et al., 2017)

The areas for AIE considerations are manifold. If AI-equipped robots become life partners or interact with elderly people who otherwise have hardly any social contact in the context of care, if algorithms spread misinformation in social networks or if no human at all can be reached in customer support - then these developments are of considerable importance for society. At the same time, it must be considered that such fears have been expressed in the case of many technological innovations, and even if some aspects such as the length of time of average daily cell phone use and the associated reduction in personal contact are certainly problematic, humans are also surprisingly adaptable. (Cf. Sharkey & Sharkey, 2012; Lazer et al., 2018)

The following sections will present relevant aspects related to developments and challenges of AIE That includes the emergence and impact of AI-based discrimination, as well as the resulting implications for the state, society, and democracy. The aim is to show which factors shape the decision-making basis of AI, which mechanisms of action are to be expected, and in which areas AI-based decisions reproduce existing social injustice.

Russell and Norvig (2003, p. 947) argue that “[a]ll AI researchers should be concerned with the ethical implications of their work”.

2.1.2. AI Ethics as a Field of Research

There is a broad variety of research in the field of AIE. It ranges from the analysis of possible implementations of ethical principles into decision routines of autonomous machines (cf. Anderson & Anderson, 2015; Etzioni & Etzioni, 2017; Yu et al., 2018) to meta-studies about AIE (cf. Vakkuri & Abrahamsson, 2018; Prates et al., 2018; Greene et al., 2019), and the comparison of ethical guidelines (Zeng et al., 2018a; Fjeld et al., 2019; Jobin et al., 2019; Hagendorff, 2020).

From theoretic empirical studies regarding the solving of trolley problems (cf. Awad, 2018) to the reflection on specific problems (cf. Eckersley, 2018). Several papers show the tremendous damage that can be done with the misuse of AI (cf. O'Neil, 2016; Brundage et al., 2018; King et al., 2012; Mehrabi et al. 2021).

Terms such as "fair machine learning", "data justice", or "discrimination-aware data mining" point to efforts by researchers across various disciplines to find issues and solutions. (Cf. Barocas & Selbst, 2016; Taylor, 2017; Veale & Binns, 2017; Binns 2018a; Dencik et al., 2019; Hoffmann, 2019) Technical issues, legal and ethical issues as well as social science aspects of the topic are increasingly coming into focus (cf. Lee & Baykal, 2017; Binns et al., 2018b; Grgić-Hlača et al., 2018a, 2018b; Lee, 2018).

Bubinger and Dineen (2021) carried out a systematic literature review in which they collected two hundred publications and used fifty of them to find actionable approaches that promote Ethical AI in libraries. They concluded that libraries have an opportunity to evaluate and minimize ethical issues of their AI-powered systems, and in that process can be leaders of ethical AI in the public sector.

2.1.3. AI-Discrimination

In addition to ethical issues, such as those of informational self-determination, moral responsibility of AI decisions, etc., a key **ethical challenge** relates to the aspect of **AI-based discrimination**. AI technologies can adopt and even reinforce human prejudices or discriminatory values and behaviors. (Cf. Heinrichs, 2021)

There are various forms of discrimination that can also be the result of automated AI-based systems. These include the denigration of minorities or disadvantaged groups of people in the linguistic and visual context, which can take the form of defamation, disparagement, or incitement to hatred by people. With the establishment of social media as important platforms for social interaction, various forms of disparagement in public and closed spaces are a common phenomenon, which is being countered with automated, but also human-made deletion measures. (Cf. Davidson et al., 2017, p. 512 f.) In addition to the partly automated deletion, social bots or chatbots, i.e., autonomously acting actors in the network, have also already attracted attention due to discriminatory posts (cf. Munger, 2017, pp. 630 f.). In the following, we will distinguish between three forms of discrimination:

Direct discrimination is the simplest form of discrimination and describes a disadvantage, e.g., in court, during police operations, at work, or at school, based on an assessment-independent personal characteristic such as gender, membership in a religion, or ethnicity (cf. Dalenberg 2018, pp. 616 ff.). This form of discrimination is also already occurring today through AI applications, such as an applicant management algorithm at Amazon that did not suggest female candidates for high-paying jobs (cf. Dastin, 2018).

Indirect discrimination refers to unfair treatment or rule that applies to everyone but has a negative impact on a particular group (cf. Doyle, 2007). This is the case, e.g., when fewer services, such as street cleaning, parks, etc., are provided in certain neighborhoods where a disproportionate number of members of a minority group live. Such indirect discrimination can also be the result of AI technologies, for example, when algorithms set prices based on customer characteristics and thus discriminate against a certain group of people. Revenue maximization through such price discrimination, which does not necessarily have to be AI-based, is already taking place today. Prices are already being adjusted to personal willingness to pay on the Internet based on usage behavior and personal data (personalized pricing) or different offers are being made based on personal data (steering) (cf. White, 2012; Hupperich et al. 2018).

Intersectional discrimination describes a form of social differentiation based on individual characteristics such as gender, ethnicity, or sexual orientation that do not occur separately but are interwoven (cf. Crenshaw, 1989, p. 141). In this context specific combinations can lead to disadvantages. For example, immigrant women may be disadvantaged even if, measured separately, there are no fundamental disadvantages against either immigrant people or women. Intersectionality is thus less easy to identify and can thus be the result of AI applications even if they are designed not to discriminate against clearly defined minorities, for example by excluding differentiation by gender or origin. (Cf. Raji & Buolamwini, 2019, pp. 5 f.)

Harms can be divided into two categories: **Representational harms** refer to the disparagement from the representation of groups, which, might negatively impact beliefs and attitudes towards the group. **Allocational harms** mean the allocation of resources or opportunities to certain groups based on the reported real-world **bias**. (Cf. Schiebinger, 2014; Crawford 2017; Olson, 2018) Alternatively, those harms can be named characterized as underrepresentation and stereotyping (cf. Dinan et al., 2020).

Bias can be detected as the cause of the harm through discrimination. Therefore, the following section is about understanding bias and its forms.

2.1.4. From Bias to Discrimination

Bias is a loaded term with to some extent overlapping, or even contending, meanings (cf. Campolo et al., 2017).

A bias generally refers to distortion effects. In psychology, it refers to attitudes or stereotypes that positively or negatively influence the perception of our environment, decisions, and actions. This influence can be unconscious (implicit bias) or conscious (explicit bias). In statistics, bias is understood as errors in data collection and processing (e.g., errors in sample selection) or the conscious or unconscious influencing of subjects. (Cf. Friedman & Nissenbaum, 1996)

In cognitive science, bias describes psychological shortcuts that can be critical to support rapid responses (cf. Tversky & Kahneman, 1973, 1974). AI research appropriated from that already existing description (cf. Rich & Gureckis, 2019; Rahwan et al., 2019). Based on it they defined bias as the discrepancy from an expected value or ideal (cf. Glymour & Herington, 2019; Shah et al., 2020). This discrepancy can occur if models rely on unintended shortcut strategies and spurious statistical cues to predict outputs (cf. Schuster et al., 2019; McCoy et al., 2019; Geirhos et al., 2020). Because of adverse

social effects that can occur based on those outcomes, bias research is not only a technical and scientific venture but especially an ethical one (cf. Bender & Friedman, 2018).

Analyzing bias is an intrinsically normative process which involves detecting what is considered as harmful behavior, to whom, and in what manner (cf. Blodgett et al., 2020; Hardmeier et al., 2021). Savoldi et al. (2021) emphasize a human centered framing of bias that puts people into the focus.

In information technology, three categories of a bias are distinguished based on the definition by Friedman and Nissenbaum. **(1) Pre-existing bias:** Often a bias that is established (pre-existing) in society is transferred to the software. This can happen explicitly, when a discriminatory feature is deliberately built in, or implicitly, when it is inadvertently built in. **(2) Technical bias:** Technical specifications - for example in sensor technology - can lead to certain groups of people being treated differently than others. This can occur through standards that do not allow certain properties to be captured, as well as the translation of human terms into mathematical models, changing the meaning. **(3) Emergent bias:** Discrimination can also arise from the interaction of software and application, such as misinterpretation of output, which often occurs with statistical values. The use of software from a certain context for use cases of a different kind also harbors this problem. Such phenomena sometimes only arise over time, for example when social patterns of action, values or processes change, but the technology does not adapt to this. (Cf. Friedman & Nissenbaum, 1996)

2.1.5. Occurring biases due to AI technology.

After introducing bias as one of the main reasons for AI-based discrimination, the following section will present particularly relevant biases in the context of AI applications: dataset bias, association bias, automation bias, interaction bias, and confirmation bias (cf. Lloyd, 2018, p. 2; Weyerer & Langer, 2020).

A given **data set bias** occurs when an AI system's data set does not adequately reflect a particular population, which can lead to biased generalizations (selection bias, e.g., for gender, sexuality, age, education, etc.). Internet data is for example not gender-neutral because women are represented in smaller numbers or differently than men in terms of content contributions (cf. Yong 2017, pp. 203 f.). From this, an algorithm could conclude that women are less able or willing to contribute. Female, young, and darker-skinned individuals are also more poorly recognized by AI-based face recognition applications than male and lighter-skinned individuals, in part because they have been trained predominantly with faces characterized by the latter features. (Cf. Klare et al. 2012, pp. 1789 ff.; Garvie 2016, pp. 8, 53; Buolamwini & Gebru 2018, pp. 1 ff.)

Association bias occurs when training data for an AI system suggest a bias that is not based on causal effects. Correlations between characteristics, events, or states are misrepresented as a causal effect relationship that is not directly present. For example, higher average salaries for men do not suggest performance. (Cf. Dastin, 2018)

Automation bias occurs when semi-autonomous systems have little human control, resulting in incorrect or undesirable outcomes. For example, in many of today's AI applications, there is a final human decision-making authority that is responsible for reviewing work practices and results of algorithms for conformity with social, moral, and cultural values and correcting them if necessary. If this human control is neglected, it can lead to decisions or actions that run counter to corresponding values and thus discriminate against certain minorities. (Cf. Kasperkevic 2015; Yeh, 2017, p. 64)

Confirmation bias can occur when information that confirms preexisting beliefs or biases is selectively perceived or preferred. Confirmation bias often occurs in the context of user profile-based recommendation systems and search engines. For example, shopping platforms recommend products that are similar to products purchased in the past or products that have already been purchased by similar customer profiles. If these are now purchased, the assumption that the recommended products were correctly recommended is confirmed. However, it is possible that the purchase decision would not have been made without the corresponding recommendation. (Cf. Chou et al., 2017)

Interaction bias can occur when an AI system learns from human communication data and infers corresponding patterns. The most prominent example of interaction bias in the context of AI is the previously described case of Microsoft's Twitter **chatbot Tay**, which will be in detail described in section 2.1.7. (Cf. Neff & Nagy, 2016, pp. 4916 ff.; Harringer, 2018, p. 261)

Filter bubbles or **echo chambers** on the Internet can be explained similarly (cf. Flaxman et al., 2016, pp. 298 ff.). It should be noted that there can be several other bias effects in the context of AI.

2.1.6. Feedback Effect

The introduction and proliferation of AI technologies is associated with the hope of establishing **greater rationality in decision-making processes** and thus counteracting subjectivity, bias, and emotional treatment of individuals. Thus, the hope is to use computer-based decision-making to counter overt or latent everyday racism and sexism. Results from computers have long been described as particularly **objective and rational**: “The ideal calculator is a computer, widely revered in part because it is incapable of subjectivity”. Hence, logical, and rule-based calculations are assumed to exclude the human burdens of biases and emotions: "the desires and biases of individuals are screened out". (Cf. Porter, 1996)

However, if technologies enjoy special credibility and, moreover, because of the contexts described above, AI results are neither comprehensible nor free of discrimination, then this may even result in a reinforcing discrimination effect, as the supposedly neutral agency of technology confirms prejudices. For example, the worldview of sexist or racist HR managers regarding the prejudice of lack of skills among women and migrants might be confirmed if the AI recruitment tool is less likely to suggest women or migrants to them for advertised management positions. (Cf. Dastin, 2018)

Discriminatory outcomes of AI applications consequently not only represent actual discrimination, but also generate a **feedback effect**. Hence, prejudices become entrenched through corresponding confirmations by supposedly neutral machines. (Cf. Weiss, 1999, pp. 29 ff.)

2.1.7. Cases of AI Bias or Discrimination

A well-known example of discrimination by AI, already mentioned in 2.1.5. as example of interaction bias, is the pilot **project Tay** by Microsoft. Tay was a self-learning chatbot with its own Twitter profile that learned how young people communicate on social media platforms. Within a few hours and high interaction on Twitter, Tay transformed into a racist, anti-Semitic, sexist, conspiratorial chatbot as its human interaction partners confronted it with corresponding content, from which it learned and recognized which posts were particularly successful in generating interactions. After just 16 hours, Microsoft was forced to take Tay offline to stop its discriminatory messages. (Cf. Neff & Nagy, 2016, pp. 4916 ff.; Harringer, 2018, p. 261)

But Tay is not the only example that can be cited. Other social bots also learn from existing entries, as a study by the Anti-Defamation League shows. 28.14% of 3060 Twitter accounts studied that sent anti-Semitic messages were identified as bot accounts. (Cf. Woolley & Joseff, 2018)

IBM's extensive AI service called Watson can also be used as an example Section. Watson processes scientific studies, Internet entries, encyclopedias, and dictionaries to provide companies and institutions with concrete answers to given problems, often in the medical field. After Watson analyzed the online encyclopedia Wikipedia and the Urban Dictionary (an online dictionary that explains colloquial terms), inappropriate language was used on several occasions, so IBM eventually set up a swear filter. (Cf. Smith, 2013)

Google's photo app provided an example for the in section 2.1.5. described **automation bias**. One of its functions automatically assigns self-generated labels to images. In 2015, this led to African people being incorrectly classified as gorillas. Unable to technically overcome the problem, Google eventually excluded the label 'gorilla' from image recognition. (Cf. Kasperkevic, 2015; Yeh, 2017, p. 64)

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool by Northpointe, Inc. It is used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. The tool uses an algorithm to estimate a value for the recidivism probability of a defendant. It was found that black people receive a worse prediction from these algorithms than white people because of their skin color. The reason for this discrimination is due to the fact that the algorithm was trained primarily with historical crime data. This data is based on statistical correlations rather than causal relationships. (Cf. Angwin et al., 2016a, 2016b)

2.1.8. Unethical use of AI

One of the more obvious ethical questionable uses of AI is in the **military** regarding cyber warfare or weaponized unmanned vehicles or drones (cf. Anderson & Waxman, 2013; Ernest et al., 2016).

Another area is the use for automated **propaganda** and disinformation campaigns. This also includes finding ways to manipulate people through manipulating algorithms on social media platforms like Facebook (cf. Lazer et al., 2018). Such platforms also enable unmonitored forms of AI experiments on society without informed consent (cf. Kramer et al., 2014).

AI can also be used for direct **social control** by governments. The Chinese scoring system is an example for that. (Cf. Engelmann et al., 2019) Another big area for possible misuse is **surveillance** – especially mass surveillance. It can be combined with face recognition or sentiment analysis to control major parts of the lives of a country’s population. (Cf. Introna & Wood, 2004; Helbing et al., 2019)

Such advanced surveillance in combination with an implemented scoring system can result in social sorting. Social sorting proposes that surveillance is not only a temporary threat to the privacy and freedom of an individual, but rather a powerful and deceitful method of generating and reinforcing social differences in the long term by assessing risks and assigning worth to individuals. (Cf. Lyon, 2005)

While such a system can be subtle, the case might be more obvious when it comes to AI **enhanced interrogation**, also known as **torture** (cf. McAllister, 2017). There are also several threats to data **privacy**. AI can be trained for the disclosure of personal traits that are private or secret. (Cf. Kosinski et al., 2013, 2015; Wang & Kosinski, 2018)

There are also unethical outcomes from the use of AI that are not related to the application itself. One of them is the environmental effect of training large language models. A research article concluded that the training of a Neural Architecture Search with a language model produces 284 tons of Carbon Dioxide equivalent (CO2e). This is comparable to five times the lifetime emissions of an average car. The training of BERT, Google's language model that enables parts of the search engine, 652 kilogram of CO2e. That is approximately as much as round trip flight from Lisbon, Portugal to Berlin, Germany. (Cf. Strubell et al., 2019, p. 4; Carbon Calculator, 2021)

Whether the in this section presented examples for the unethical use of AI are truly, by their own nature, unethical is open to a philosophical debate, and depends on the specific context.

2.1.9. Deepfakes

AI can already be used to fake a whole person with a synthetic voice and deepfake visuals. In that way technology can be used for personality theft, fraud, but also to destabilize whole societies. (Cf. Bendel, 2017; Güera & Delp, 2018)

Deepfakes are increasingly realistic-looking photos, audios, or videos in which people are placed in new contexts or words are put into their mouths with the help of AI technologies that were never said that way. The technology certainly opens new possibilities for artists, for digital visualizations in schools or museums, and helps in medical research. At the same time, however, deepfakes entail considerable dangers, as the international study now presented for the Scientific Technology Options Assessment Committee (STOA) of the European Parliament shows. The technology can be misused to spread fake news and disinformation very effectively, e.g., fake audio documents could be used to influence or discredit legal processes and ultimately threaten the judicial system. It would also be possible, for example, to use a fake video not only to harm a politician personally, but also to influence her party's election chances and, ultimately, to damage trust in democratic institutions. It is new generation of digitally manipulated media content that has become cheaper and easier to produce in recent years and, above all, can look deceptively real. (Cf. STOA, 2021) While AI can be used for the creation and spread of fake news, it should be noted that it can also effectively be used for the fight against fake news by detecting it (cf. Nakov et al., 2021).

The researchers from Germany, the Netherlands and the Czech Republic propose concrete solutions. Due to rapid technological progress, they say, one should not limit oneself to regulations on technology development. To be able to manipulate public opinion, fakes not only have to be produced, but above all they must be disseminated. In regulating how to deal with deepfakes, we therefore need to start with Internet platforms and media companies first and foremost. However, AI-powered technologies for deepfakes are unlikely to be eliminated even in this way. On the contrary, the researchers are convinced that individuals and societies will be increasingly confronted with visual disinformation in the future. It is therefore essential to be even more critical of such content in the future and to further develop skills that help to critically question the credibility of media content. In addition to ITAS, the Fraunhofer Institute for Systems and Innovation Research was involved in the study on the German side, the Rathenau Institute as project coordinator in the Netherlands, and the Technology Centre CAS in the Czech Republic. (Cf. STOA, 2021)

2.1.10. Mitigation and Key Findings

Safeguards exist that can prevent potential harm and discrimination. There are already a variety of initiatives to develop remedies against AI-related bias and discrimination. Technical solutions are being developed to identify and counter AI bias and discrimination. Analysis tools aim to help developers understand how an AI system makes its decisions. **Traceability** of AI-based decisions is one of the biggest challenges, as without it, one can never be sure whether a decision is legitimate from an ethical perspective or not (black box challenge). In addition, other tools can be used to anonymize personal characteristics to ensure compliance with **privacy** policies. (Cf. Bellamy et al., 2018; IBM, 2018a, 2018b; Accenture, 2018; Google, 2019; Brighter.ai, 2021) We will further elaborate on those safeguards in section 4.1.4. and 4.3.4.

Several researchers reflected on the idea that **ethical guidelines** serve as a basis for ethical decision-making for software engineers. The finding was that the effectiveness of guidelines or ethical codes is almost zero and that they do not change the behavior of professionals from the tech community. No statistically substantial disparity in the replies was found across individuals who did and did not see the code of ethics, either for students or for professionals. (Cf. McNamara et al., 2018) Boddington (2017, p.56) concluded that **ethical considerations are mainly used for public relations purposes**.

The mitigation needs involvement and collaboration by **governments**, **academic research**, **NGOs**, and the **industry**. Those entities are defined as factors to be compared for the Case Study. The results of AI applications should be treated with caution regarding their potentially discriminatory output. Already today, the selection of content often generated by AI applications in search queries and social media walls is strongly biased towards the user's own profile, which influences and thus potentially distorts the perception of reality. The outcome is an echo chamber which can lead to polarization of individuals, and division within societies. (Cf. Passe et al., 2018; Törnberg, 2018; Baumann et al., 2019)

This bias and the associated perpetuation of existing prejudices can reinforce the polarization of a heterogeneous society and thus represents one of the many dangers of the digital transformation of the state (i.e., political order and public institutions) and society (i.e., the totality of people in the state and corresponding social fabric and values). However, to assess the application of AI and its potentially discriminatory failures, it should be evaluated whether not applying AI creates more benefits for the state and society: "any fair assessment of algorithms must be made against their alternative." (Cf. Thierer et al., 2017, p. 37)

Methods have been previously developed to debias language models. They involve pre-processing the training datasets in specific ways (cf. Lu et al., 2020) or adjusting the training algorithm (cf. Qian et al., 2019). More research is needed to debias large language models because such models are starting to be used in various real-world tasks. Model Debiasing: Gender tagging, Adding context, Debiased word embeddings. Model de-biasing through **explainability** or **interpretability** techniques to identify spurious statistical cues (cf. Belinkov et al., 2020). Due to the enormous application areas within AI, it is impossible to make general ethical statements; however, ethical codes can and must be established to regulate the mostly industrial developments. It becomes clear that it is not just about individual developments of artificially intelligent programs and the question of whether their behavior is moral or not in individual cases, but also about the question of **responsibility**: Who decides on the development, production, and use of programs? And how does one deal with malfunctions or inhumane consequences?

3. CASE STUDY DESIGN

3.1. CASE STUDY METHODOLOGY

The qualitative case study design is a suitable method to investigate and better understand complex phenomena in their context. There is a variety of scientific theoretical literature regarding the case study methodology (cf. Yin, 2003; George & Bennett, 2005; Gerring, 2016).

The case study design is widely used in qualitative social science research. Many terms exist, such as "single case study" or "case reconstruction," which are often used synonymously with Case Study Design. However, the term "qualitative case studies" per se does not say anything about the methodology but forms a framework term for using a wide range of research methods. Qualitative case studies have their origins in the basic idea of qualitative research . Like other qualitative research methods, they follow the constructivist paradigm's principles, which states that truth is relative and dependent on one's perspective (cf. Baxter & Jack, 2008; Simons, 2008). However, the distinctive feature of qualitative case studies is that the focus is on explaining and examining the context of a phenomenon and its influence (cf. Yin, 2009; Cresswell & Poth, 2016). This makes clear what case studies are fundamentally about: enabling a holistic understanding of a phenomenon.

The focus is on an in-depth and broad examination of a case or cases to obtain as complete an overview as possible of the entity or phenomenon under investigation. In this context, the concepts of "totality" and "different perspectives" play an essential role. To be able to grasp a case in its entirety and its complexity, a variety of different data sources is often required. (Cf. Yin 2009, p. 101; Webb 2014).

There are two critical things about why a case becomes a case for a case study. The first essential element is the subject - the practical unit of a case. This can be, for example, a person, a place, a small group of people, and organizations, families, or social groups or systems. In addition, a case must also have fundamental or specific characteristics related to the research interest and the research question and thus form the theoretical and scientific basis of a case - the analytical framework. (Cf. Webb 2011)

There are many variations of the qualitative case study in the literature. For example, while Gillham (2000) and Yin (2009) distinguish between single cases and multiple cases, Stake (1995) speaks of collective, instrumental, and intrinsic cases, and Yin (2011) of descriptive, explanatory, exploratory, and evaluative case studies.

Ultimately, there is hardly any unified direction in the literature to describe the characteristics of case studies. This is also the reason for the problems around qualitative case study design. This method is described, handled, and used very differently in the literature and textbooks. In published empirical work based on case studies, there is often no precise description of the characteristic features, the designs used in the studies and the methodological procedure.

3.2. CASE STUDY APPROACH DESIGN

The design of the case study approach stands on the preceding considerations from the literature review and the case study methodology. Based on the literature review six factors are defined to compare the different markets as cases. Those factors aim to be an objective measurement for the comparison. The case analysis is conducted based on scientific papers, official press releases by non-governmental organizations, governments, academia, and companies.

Three markets are analyzed and compared based on the defined factors. Those are by name the USA, EU, and China. Each is described superficially in the beginning of each section, including the reasoning why this market was chosen for this case study. The aim of this approach is to gain an in-depth, multi-faceted understanding of the complex issues of AIE in its real-life context. The analysis of the factors is followed by a discussion regarding the finding from 1. to 5. of each case. The discussion questions the tradeoff between innovation and ethics and revisits the cases' limitations. The findings during throughout the literature review led to the following factors for the Case Study:

1. **Government**. It plays the biggest role when it comes to legislation and regulation regarding AIE. This includes highlighting the position towards military and surveillance use of AI.
2. **NGOs & NPOs**. Organizations that are neither governmental nor part of the industry. Namely, e.g., Universities, Civil-Rights groups, Thinktanks. It also includes relevant individuals that are not directly affiliated with the government or industry.
3. **Industry**. As shown throughout the LR, companies, e.g., Microsoft, Google, are engaged with AIE. They are the ones that benefit monetarily through products and services. They are also possibly the ones struggling the most with regulations.
4. **Guidelines, Frameworks, Principles**. Those are often a collective work between two or all before-defined factors. This factor will be a part of the three factors.

The Case Study should therefore analyze in more detail how those factors are engaged in all three markets. Based on the findings of the literature review, the Case Study should furthermore also emphasize the following aspects in each market regarding AIE: Data quality and quantity, control and accountability, transparency, military, propaganda, surveillance, social control, privacy (specifically data privacy), black box challenge, deepfakes, enhanced interrogation or torture, traceability, de-biasing, explainability or interpretability, responsibility.

All those aspects have been named throughout the literature review. Thus far, the conducted research has not laid any substantial emphasis on markets. This will change on the following pages. The author will try to be objective as possible while doing so, but a certain bias in favor of liberal western democracies is undeniable. Besides using scientific literature, there is also popular science media, e.g., wired, and statements or documents by governmental organizations, e.g., DoD or European Commission, used for the Case Study. While introducing the different markets, the **Government AI Readiness Index 2020** will be used which compares 172 countries with ten factors. "The index measures governments' readiness to implement AI in the delivery of public services to their citizens; it looks at the capabilities and enabling factors required for a government to be ready for AI implementation, but it does not measure the implementation itself." (Cf. Shearer et al., 2020) Every market introduction consists of general economic factors and data related to AI. This includes the number of AI startups to emphasize the relevance for AIE.

4. CASE STUDY EXECUTION

4.1. USA

4.1.1. Introductory Considerations

The USA ranks highest in the AI Readiness Index. In the Governance & Ethics category it scores 92.66 out of 100 which is the highest score in the compared countries. (Cf. Shearer et al., 2020, p. 27) The US government is intending to spend two billion USD for military AI projects between 2018 and 2023 (cf. Fryer-Biggs, 2018). The US has 10.099 AI startups by June 2021. (Tracxn, 2021a) They have a well-known AI industry with global players, e.g., Google, Microsoft, and Nvidia (cf. Botha, 2019).

The Pew Research Center conducted a survey with 4,135 adults in the US in 2017. One of the findings was that numerous Americans anticipate significant impacts from a variety of automation technologies during their lifetimes. Those range from the widespread adoption of autonomous vehicles to the replacement of entire job categories with machine-based workers. (Cf. Smith & Anderson, 2017)

A report from 2019 based on the survey of 2,000 US adults during 2018 concluded that Americans are overall worried about a possible AI catastrophe. A substantial majority of Americans agreed that AI systems need careful supervision. The respondents said that an AI apocalypse is less likely to happen the failure to address climate change but ranked it as more catastrophic if it occurs. (Cf. Zhang & Dafoe, 2019)

The same authors conducted a survey in 2020 which revealed that while Americans consider AI governance challenges to have a level of importance, they do not necessarily trust the actors that have the power to manage and develop the technology to act in the interest of the public. The distrust does not necessarily predict opposition to AI development. (Cf. Zhang & Dafoe, 2020)

The Tencent Research Institute concluded in 2017 that the US at this point was the country that had published the biggest amount of policy reports and strategies regarding AI governance and ethics. They concluded that the US was undoubtedly the forerunner in the field of AI research and that every move necessarily would have global effects. (Cf. TRI, 2017)

4.1.2. Government

In 2019, the **White House** issued an Executive Order i.e., “Maintaining American Leadership in Artificial Intelligence”. The order implied that the **National Institute of Standards and Technology (NIST)** has to develop a plan to set up technical standards for reliable, robust, and **trustworthy AI** systems.” (Cf. E.O., 2019) In 2020, the Guidance for Regulation of Artificial Intelligence Applications was created because of the issued Executive Order. Principles for the **Stewardship of AI Applications** by the White House **Office of Science and Technology Policy (OSTP)** in 2020. (Cf. Vought, 2020, pp. 3-7)

Another related Executive Order was issued by the **White House** in December 2020, i.e., Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. This order defined nine principles for the use of AI in government. (Cf. E.O., 2020). Eric Lander, science adviser to the president and director of the White House OSTP, and Alondra Nelson, deputy director for science and society at the White House Office of OSTP, called for the development of a **Bill of Rights** for Americans in a world powered by AI in August 2021. (Lander & Nelson, 2021)

In October 2021, the OSTP made the first step for a **Bill of Rights** to limit AI harms by launching a fact-finding mission with a call for interested parties to participate in the assessment of AI-enabled biometric technologies public and private sectors (cf. Murphy, 2021).

In 2019, the **Department of Defense (DoD)** in collaboration with the Defense Innovation Board (DIB) developed AIE Principles. With those principles, the department set the goal that its use of AI is: (1) Responsible, (2) Equitable, (3), Traceable, (4) Reliable, and (5) Governable. To reach those goals, the DoD made twelve recommendations, of which one is the establishment of a department wide AI Steering Committee for the oversight and execution of the defined AI principles. The authors ended the document mentioning that those principles and recommendations shall not restrict the Department's capabilities and the DoD is a deeply ethical organization. (Cf. DoD, 2019) The five principles have been officially adopted by the DoD in February 2020 (cf. DoD, 2020).

The **National Security Commission on Artificial Intelligence (NSCAI)** reported in November 2019 that there is a risk that AI-enabled systems could track and attack previously invulnerable military positions, undermining global strategic stability and nuclear deterrence. States could be tempted to behave more aggressively as a result, which could increase incentives for a first strike. (Cf. NSCAI, 2019, p. 11) The report proposes agreements among the United States, Russia, China, and other nations to seek a ban on the launch of nuclear weapons authorized or triggered by AI systems (cf. NSCAI, 2019, p. 46).

In 2021, the NSCAI released its final report. The report advocates for several approaches in relation to AIE and proposes recommendations for the US and its governmental agencies. There need to be ethical constraints regarding where and when AI can appropriately be used within a human-AI team in each context. It calls for feasible metrics and testing methodologies to enable the evaluation of requirements for principles of ethical, responsible, and trustworthy AI. They conclude that ethical standards for the development of AI are lagging behind the technology itself. The team around Schmidt proposes a fund that invests in the development of AI applications that bear with ethical norms and democratic values. They identified stricter export control rules as a measurement to promote the responsible and ethical use of AI among US companies, which could set standards for the global industry, and therefore counter civil rights abuses. (Cf. Schmidt et al., 2021)

In October 2021, the Director of the **Artificial Intelligence and Technology Office (AITO)** at the Department of Energy (DoE) spoke about the **Agency's Plan to Advance Trustworthy AI**. The Director and her team the AI Risk Management Playbook (AI RMP), a system only available to DoE users thus far. The AI RMP offers more than a hundred unique risk and mitigation techniques. One notable example is the introduction of a lifecycle process for trustworthy AI. It is a comprehensive system that directly relates to the Executive Orders mentioned before. They work closely with the White House and plan on working with the NIST. (Cf. Isom, 2021; AITO, 2021)

The **National Institute of Standards and Technology (NIST)** AI Risk Management Framework is currently in development and scheduled to be published in 2023. They develop the framework in a transparent process and encourage the participation of the public and private sectors. The NIST is part of the U.S. Department of Commerce. (Cf. NIST, 2021a) Microsoft is one of the companies participating by providing feedback and insights into their own handling of AI risks (cf. NIST, 2021b). So far, they have received a total of 106 comments in response to their open request for information (cf. NIST, 2021c).

The **Defense Advanced Research Projects Agency (DARPA)** aims to tackle the problem of **explainability**. With their technology for explainable AI, they aim to enable users to understand, effectively manage, and appropriately trust the rising generation of AI systems. The objective for doing so is to generate more explainable models, while maintaining a high level of learning performance respectively prediction accuracy. (Cf. Turek, 2021).

4.1.3. NGOs & NPOs

In 2016, **Partnership on AI (PAI)**, an NPO with the vision of a future in which AI empowers humanity by contributing to a more just, equitable, and prosperous world, released six pillars and eight tenets. One of those pillars is the pursuit of fair, transparent, and accountable AI. (Cf. PAI, 2016)

The PAI has currently 95 global partners with the majority of 64 coming from the US. Partners from the US range from big corporations like Amazon to media outlets like The New York Times, and civil rights organizations like the American Civil Liberties Union (ACLU). (Cf. PAI, 2021)

The **Asilomar Conference on Beneficial AI** by the **Future of Life Institute (FLI)** in 2017. The conference resulted in the 23 Asilomar AI Principles. Elon Musk participated. There have been 37 researchers and 45 scientific publications funded by the Future of Life Institute's AI Safety Research program as of 2018. 3462 robotics/AI researchers signed an FLI open letter to ban autonomous weapons. (Cf. Ding, 2018, p. 30)

The OpenAI Charter by **OpenAI** was established in 2018. They claim that their fiduciary duty is to humanity with the objective of enabling AI deployment for the benefit of all, to no harm humanity or unduly concentrate power. (Cf. OpenAI, 2018)

The central backers of **OpenAI** are, among others, Elon Musk, Peter Thiel, and Microsoft. The objective of OpenAI is to develop and commercialize open-source artificial intelligence in a way that benefits society, not harms it. The organization enables "free collaboration" with other institutions and researchers by making its patents and research results available to the public. (Cf. Gershgorn 2015; Lewontin 2015)

Other notable NGOs & NPOs: **AI Policy Principles** by the **Information Technology Industry Council** (ITI) in 2017. Guiding Principles and Recommendations by the **Internet Society** in 2017. Principles for the Governance of AI by **The Future Society** in 2017. Principles for Algorithmic Transparency and Accountability by the **ACM US Public Policy Council** (USACM) in 2017. Three Rules for Artificial Intelligence Systems by the **Allen Institute for Artificial Intelligence** in 2017. Universal Guidelines for Artificial Intelligence by **The Public Voice coalition** in 2018. **The Stanford Human-Centered AI Initiative** (HAI) by the Stanford University in 2018. Seeking Ground Rules for A.I.: The Recommendations by **New Work Summit** in 2019.

The following illustration shows the number of peer-reviewed AI publications in the US by institutional affiliation from 2000 to 2019. (Zhang et al., 2021, p. 22)

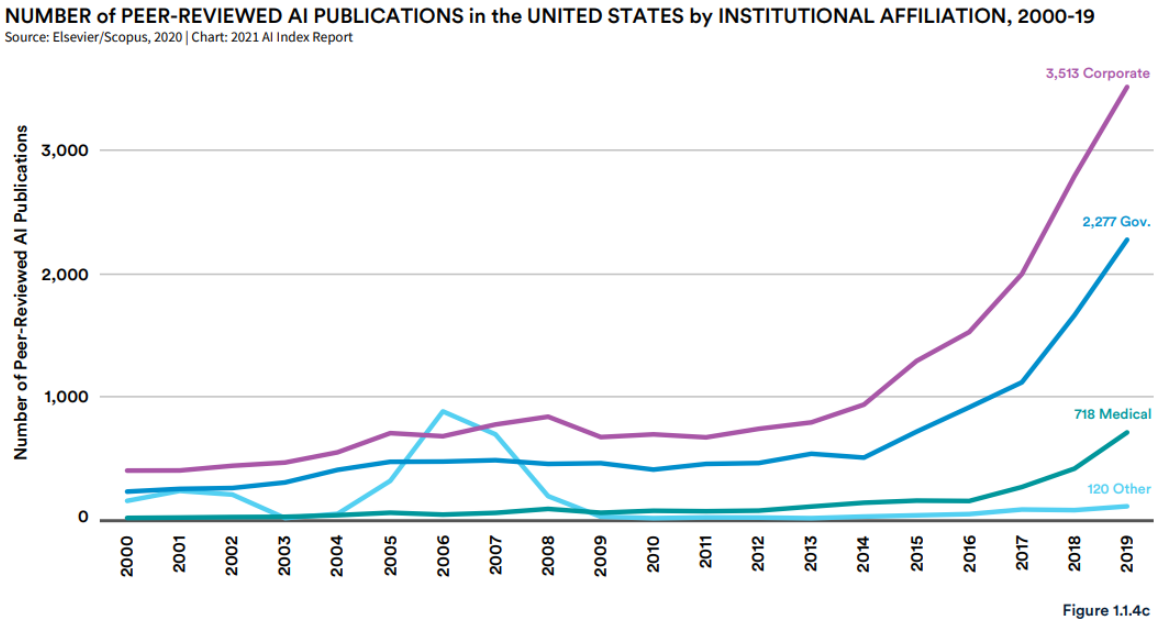

Figure 6 - AI publications in the US by institutional affiliation from 2000 to 2019 (Zhang et al., 2021, p. 22)

4.1.4. Industry

Google published 2018 Our Principles by Google. They defined seven principles, which are: (1) Be socially beneficial. (2) Avoid creating or reinforcing unfair bias. (3) Be built and tested for safety. (4) Be accountable to people. (5) Incorporate privacy design principles. (6) Uphold high standards of scientific excellence. (7) Be made available for uses that accord with these principles. In addition to those principles, they also state that they will not deploy AI application for four areas, which are: (1) Technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks and will incorporate appropriate safety constraints. (2) Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people. (3) Technologies that gather or use information for surveillance violating internationally accepted norms. (4) Technologies whose purpose contravenes widely accepted principles of international law and human rights. (Cf. Google, 2018)

Deepmind, a Google subsidiary, recently was involved in the NHS health data scandal. They allegedly received personal records of 1.6 million patients without the patients complying to it. The company argued that the information was used for the development of a potentially lifesaving application but apologized for the usage without consent. (Cf. Daws, 2021).

Sundar Pichai, CEO Alphabet, believes that AI will have a fundamental impact on the development of humanity similar to that of fire or electricity. Pichai is not alone in this conviction in the AI industry. AI is expected to make progress where humanity reaches its limits and solve its fundamental problems: Climate change, social injustice, disease. (Cf. Pichai, 2018)

Eric Schmidt, former executive chairmen of Alphabet and current chairmen of the NSCAI, believes that "China could soon replace the U.S. as the world's AI superpower", arguing that "there are serious military implications to consider" and urged "President Biden to reject calls for a global ban on highly controversial AI-powered autonomous weapons". He thinks that "China and Russia are unlikely to keep to any treaty they sign." (Schmidt, 2021a, 2021b)

Google is pursuing a research approach to address **traceability** and take on the **black box challenge** called Testing with **Concept Activation Vectors** (TCAV) to identify and counter AI bias and discrimination. In doing so, the TCAV analysis tool aims to help developers understand how an AI system makes its decisions. of AI-based decisions is one of the biggest challenges, as without it, one can never be sure whether a decision is legitimate from an ethical perspective or not. (Cf. Google, 2019).

Project Maven, previously mentioned in section 1.1., the military AI program too unethical for Google, was taken over by Palantir in 2019 (cf. Greene, 2019).

10 AI Laws by **Microsoft** in 2016, which are: (1) AI must be designed to assist humanity. (2) AI must be transparent. (3) AI must maximize efficiencies without destroying the dignity of people. (4) AI must be designed for intelligent privacy. (5) AI needs algorithmic accountability so humans can undo unintended harm. (6) AI must guard against bias. (7) It's critical for humans to have empathy. (8) It's critical for humans to have education. (9) The need for human creativity won't change. (10) A human has to be ultimately accountable for the outcome of a computer-generated diagnosis or decision. (Cf. Nadella, 2016)

Microsoft's declared that for them questions such as "Why are we building this AI system?" or "Is the AI technology at the core of this system ready for this application?" are important. The answers to such questions help determine whether an AI solution will meet with the necessary acceptance from in-house users and customers. Microsoft engages in sharing experience, providing open-source tools such as Fairlearn or InterpretML, and with the creation of the "Responsible AI Learning Lab", which is a workshop that guides participants through real application scenarios of AI. (cf. DISER, 2021; Langkabel, 2021).

Principles for the Cognitive Era by **IBM** in 2017, and Principles for Trust and Transparency, and Everyday Ethics for Artificial Intelligence: Five Areas of Ethical Focus in 2018. The later one substantially influenced the European AI Alliance. Five factors are: (1) Accountability: AI designers and developers are responsible for considering AI design, development, decision processes, and outcomes. (2) Value Alignment: AI should be designed with consideration of the norms and values of the user group. (3) Explainability: AI should be designed for humans to easily perceive, detect, and understand its decision process. (4) User Data Rights: AI should be designed to protect user data and preserve the user's power over access and uses. (5) Fairness: AI should be designed to minimize bias and promote inclusive representation. (Cf. IBM, 2017; Rossi, 2018) IBM also created one of the most comprehensive technical approaches to solving AI-related biases and discriminations with the **AI Fairness 360** Toolkit (cf. Bellamy et al., 2018; IBM, 2018a, 2018b).

Attempts to develop remedies against AI-related bias and discrimination are being made in the IT industry, which faces a critical clientele. Consulting firm **Accenture** has introduced AI testing services based on a teach-and-test methodology to help companies to deploy AI systems free of bias and discriminatory content (cf. Accenture, 2018).

Accenture has identified AEI as a business opportunity by acknowledging the growing demand for incorporating ethical considerations into AI products and services. They argue that there is not much guidance on incorporating ethical consideration. To fill this gap, they partnered with Northeastern University to explore the development of practical and well-functioning data and AI ethics committees. (Cf. Sandler et al, 2019)

Other notable AEI principles, initiatives or developments in the industry: AI public policy principles by **Intel** in 2017 (cf. Intel, 2017). **Unity's** Guiding Principles for Ethical AI by Unity Technologies in 2018 (cf. Unity, 2018). **GE Healthcare** AI principles by GE Healthcare in 2018 (cf Bigelow, 2018). The National Science Foundation (NSF) Program on Fairness in Artificial Intelligence in Collaboration with **Amazon** (cf. Leen et al., 2021). **Facebook** trained its AI to block violent live streams after Christchurch attacks (cf. Sabbagh, 2021). **Amazon** dump secret AI recruiting tool that showed bias against women (cf. Dastin, 2018).

4.2. CHINA

4.2.1. Introductory Considerations

The People's Republic of China, commonly and in this thesis referred to as China, is a country in East Asia. With more than 1.4 billion inhabitants, China represents the most populous and the third largest country in the world in terms of total land area. (Carter, 2021; Worldometer, 2021; WSP, 2021) China is one of biggest economies with an estimated GDP for 2021 of $16.64 trillion by the International Monetary Fund (cf. IMF, 2021). There are 1,513 AI startups in China as of June 2021 (cf. Tracxn, 2021b).

China ranks 19 in the **AI Readiness Index** which compares 172 countries The score for Governance & Ethics is 85.58 which is one of the highest, but lower than Singapore, US, and Canada. On the sub-Index for **responsible use of AI** China ranks 34 out of 34. This sub-index measures 9 indicators across 4 dimensions: Inclusivity, Accountability, Transparency, and Privacy. (Cf. Shearer et al., 2020, p. 21 f., 116)

China is generally recognized as one of the greatest AI powers in the world (cf. Shearer et al., 2020). They are known in the AI industry for its large label companies which manually prepare data sets for supervised learning with large amounts of data (cf. Yuan, 2018). AI may be the first technology domain in which China becomes the global trendsetter (cf. Ding, 2018).

In July 2017, the New Generation Artificial Intelligence Development Plan by China's State Council aims to be world leader by 2030 to monetize AI into a 150 billion dollars industry. China wants to emerge as the driving force in defining ethical norms and standards for AI. (Cf. State Council, 2017)

In 2018, Ding researched regarding the myth that "there is little to no discussion of issues of AI ethics and safety in China". He concluded that there is "Substantive discussions about AI safety and ethics are emerging in China. (Cf. Ding, 2018)

China is one of the leading countries in AI. Therefore, their approach to AI regulation will play an essential role in managing the distinct risks of AI technology. This includes risk scenarios involving the misuse of AI and AGI as outlined by experts in recent years. (Cf. Bostrom, 2014; Brundage et al., 2018)

The public is generally not seen as the decisive force in China's AI development, but Chinese citizens are advocating for ethical constraints in some instances. There is a widespread perception in the West that Chinese people are particularly trusting of new technologies which does not appear to be true. There is growing debate, awareness, and sometimes opposition related to the risks that emerge with AI. This has led to corporate self-regulation and policy changes in some cases. Chinese citizens overall care about the protection of their personal information. (Cf. Arcesati, 2021)

The occurrence of surveillance technologies that are powered by AI worries citizens. Over 70 percent of respondents in a survey conducted in 2019 expressed concerns on the subject of privacy violations in the implementation of facial recognition systems. (Cf. Lin, 2019)

Several civil lawsuits have drawn attention to citizens' privacy concerns with China's growing use of facial recognition in public spaces. In 2019 in Hangzhou the first lawsuit was filed against a wildlife park for introducing an access-control system that utilizes facial recognition. The individual won the historic case, which ignited intense online debates about the excessive collection of facial data. Various other high-level-profile lawsuits followed. (Cf. Zeng et al., 2019)

A separate noteworthy public backlash occurred in reaction to the rise of in section 2.1.9. described deepfakes in China. The backlash occurred in 2019 when the release of the face-swap/deepfake app ZAO caused almost-instant outrage among its users over data privacy. (Cf. Porter, 2019)

4.2.2. Government

In 2020, the **National New Generation Artificial Intelligence Governance Committee** released the **New Generation AI Governance Principles** – Developing Responsible AI. They defined eight governance principles in their first document regarding AIE, which are: Harmony and friendship, Fairness and justice, Inclusive and sharing, Respect for privacy, Safety and controllability, Shared responsibility, Open collaboration, and Agile governance. (Cf. CIF, 2020)

China's **New Generation of Artificial Intelligence Development Plan** Implementation Office to coordinate inter-ministry implementation. The Government has also set up a New Generation AI Strategic Advisory Committee to drive forward the adoption of AI. China scores reasonably well in the Governance and Ethics dimension and has published its own set of Governance Principles to regulate the use of AI. (Cf. China Daily, 2019)

However, China's vision of AI to promote **social harmony** has often led to intensive social intensive social surveillance. A 2018 survey revealed that over 75% of respondents felt that AI is a threat to **privacy**. (Cf. Hersey, 2018) The first time the Chinese government outlined an agenda for AI safety measures occurred in the State Council's New Generation Artificial Intelligence Development Plan in 2017. The document announced a roadmap on which by 2025, China will have initially established AI laws and regulations, ethical norms, and beginnings of AI security assessment and control capabilities. Furthermore, based on the roadmap China will have developed ethical norms and a policy system, as well as more comprehensive AI laws and regulations by 2030. (Cf. State Council, 2017)

The development plan aims to create guidelines like an ethical framework for codes of conduct for people in AI product R&D,and in the design of human-machine collaboration (cf. State Council, 2017

There were no further specifics given at that point, which resulted in receptions of the document calling it opaque regarding ethical AI research (cf. The Economist, 2017). Chinas President Xi Jinping called in 2018 for the "healthy development" of AI based on institutional mechanism, laws, regulations, and ethics. It's an important driving force for the new round of scientific and technological revolution and industrial transformation. (Cf. Lifang, 2018)

2018 AI Standardization White Paper by the **Chinese Electronics Standards Institute (CESI)** issued three main ethical considerations for AI: (1) humans interest; (2) liability; (3) consistency of rights and responsibilities. The paper discusses privacy, safety, and ethical issues, and echoes the governments aim to introduce technical standardization as a tool for global and domestic AI governance endeavors. (Cf. CESI, 2018)

The **Ministry of Science and Technology (MOST)** formulated in 2019 the most official Chinese governance principles regarding AIE to date. They issued eight principles for responsible AI, created by a dedicated expert group. Those principles are (1) harmony and friendliness; (2) fairness and justice; (3) inclusivity and sharing; (4) respect for privacy; (5) safety and controllability; (6) shared responsibility; (7) open collaboration; (8) agile governance. (Cf. MOST, 2019)

The committee responsible for the MOST principles is comprised of experts from esteemed universities, AI companies, and the Chinese Academy of Science (cf. BAAI, 2019). MOST is trying to drive implementation at the local level. The encouragement by MOST aims at municipal governments to step up, which has resulted in AI pilot zones. The MOST established AI pilot zones that develop initiatives to implement principles on the ground. Fifteen of such zones have been announced so far, and five more a planned to be established until 2023. (MOST, 2019)

Among those are: Tianjin, with an explicit mandate to build an AI Governance Platform, and governance technologies. Shanghai, with proposals on the collaborative implementation of AI Governance Principles, and the establishment of AI governance conferences. Hangzhou, with the focus on security frameworks and technical standards for healthy AI development, e.g., for smart city governance and autonomous vehicles. (Cf. Jia, 2020b; Wei, 2020)

Recently China tightened restrictions regarding the use of AI for medical judgments, i.e., that AI software should not be used as a substitute for conclusions – which must be conducted by a registered doctor (cf. Wei, 2021). In 2019, Beijing hosted the first meeting of an influential standardization committee for AI (cf. Ding, 2018).

Facial recognition regulation has received increasing attention from top lawmakers in China. The recently enacted Civil Code and the abovementioned personal information regulations both tighten restrictions over the collection of biometric data. Standard-setting authorities more recently released a draft for dedicated national data security standards especially for facial recognition data. (Cf. Wang, 2021) Various cities have also considered their own regulations or have already introduced some to restrict the use of facial recognition. Several are penalizing companies for data privacy violations. (Cf. Feng, 2021)

Chinese authorities were fast in taking preliminary measures to regulate the use and distribution of deepfakes. In reaction to before mentioned issue with the ZAO App one released policy document specified the requirement of online information service providers to review and label all audiovisual content that is created by using innovative technologies such as deep learning. (Cf. Zhong, 2019)

Following regulations furthermore forbid the use of deep learning to transmit, publish, or create fake news (cf. Arcesati, 2021). Those cases can show that civil society influences corporate actions and government regulations related to AIE to a certain degree. Their ability to do so however is in the end constrained by China's political system. The data protection regime in China provides the government with unrestricted power to collect and use their citizens' data for its invasive law enforcement and public security activities. (Cf. Horsley, 2021)

China has anticipated a ban on lethal autonomous weapon systems (**LAWS**). The proposed ban from defines LAWS so narrowly that it would probably not constrain China’s development or use of these weapons even if the international community accepted it. Hard to say whether Chinas concerns about **moral responsibility** and human dignity are genuine. In part because China is less sensitive to some other ethical concerns, e.g., their residents’ rights to privacy. (Cf. Morgan, 2020, p. 123)

There is reason to believe though that Beijing does genuinely care about the strategic and operational risks caused by **military AI**. No political or military leader wants lethal weapons that can be hacked, or that may perhaps show unpredictable emergent behaviors. Nor does any state leader want their military commanders counseled by decision support systems that might propose actions that are insensitive to escalation levels and thus risk **stability** which could lead to escalation in war. Such concerns might be even greater in China than in some other countries, because of its political and strategic culture, which put emphasis on **centralized control**. (Cf. Morgan, 2020, p. 123)

4.2.3. NGOs & NPOs

The Carnegie–Tsinghua Center for Global Policy is currently listed as the only Chinese partner for the in section 4.1.3. mentioned PAI. Digital Asia Hub and the Centre for Artificial Research are also Chinese partners when taken Honk Kong into account. (Cf. PAI, 2021)

China's top Al research institute emerged in 2018 as the **Beijing Academy of Artificial Intelligence (BAAI)**. The institute is a hub for multistakeholder and international collaboration. Zeng Yi leads the BAAI’s research center, which has the objective to investigate Al governance and ethics. A 2020 published study by BAAI in collaboration with researchers at Cambridge University aims to promote international discourse. The study encourages academia to play a bigger part in overcoming cultural difficulties for a broader cooperation on AI governance and ethics. (Cf. ÓhÉigeartaigh et al., 2020)

In 2019, the Beijing AI Principles by the BAAI have been founded by several Chinese Universities, including Peking University and Tsinghua University. It is backed by the Beijing municipal government and **MOST**. (Cf. BAAI, 2019)

None of the research for the global **Asilomar Conference on Beneficial AI** by the **FLI** introduced in section 4.1.3., was conducted at a Chinese institution. Of the 3462 researchers who signed the open letter to ban autonomous weapons, only three were based at Chinese institutions. All of those were affiliated with the Chinese University of Hong Kong. (Cf. Ding, 2018) Out of more than 150 attendees, only Andrew Ng was working at a Chinese company at the time, but shortly afterwards resigned from

his role at Baidu. Ng, in the position of Chief Scientist, created the company's AI Group which now consists of several thousand people. (Cf. Ng, 2017; Mozur, 2017)

Renowned scholars have meanwhile in addition to regulation advocated for the use of technological processes and measures, i.e., ethics by design to ensure the **responsible** use of biometric data. (Cf. Zeng et al., 2019)

Joint Pledge on AI Industry Self-Discipline launched by the Artificial Intelligence Industry Alliance (AIIA), a group of technology companies and an association of universities. It's led by the China Academy of Information and Communication Technology (CAICT) and the Ministry of Industry and Information Technology (MIIT). The MIIT is the top government-affiliated think tank for technology policy issues. (Cf. Webster, 2019)

Beijing AI Principles and the Joint Pledge aim for action-oriented and applicable goals, and initiatives that facilitate the AI development, from R&D to commercialization, lifecycle of system, ensuring that it is beneficial for society (cf. Gal, 2020, p. 53).

Jeffrey Ding in collaboration with Brian Tse concludes that China has overall a low level of engagement with Western institutions and countries on discussions regarding AI safety across academic, public, and private sectors. A variety of Chinese AI researchers are translating the IEEE's Ethically Aligned Design report, which is part of the Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. (Cf. Ding, 2018)

There are diverse perspectives on AIE and safety in the AI community in China. The **China Academy for Information and Communications Technology (CAICT)** hosted an event in 2017 on the unique challenges AI poses for governance and law. Attendees included the dean of Tsinghua University law school, Weixing Shen, Tencent's Chief Research Officer, Guobin Li, president of the Beijing Research Institute for Communication Law, Si Xiao, and representatives from the Supreme People's Court, the highest trial organ in the country. It appears that participants offered robust and, often diverge, views on how to govern and regulate AI. Dean Shen, for instance, defined the ongoing AI developments as an unchangeable social trend that should be embraced rather than exceedingly worried over. In contrast Li argued that scholars should proactively address the policy and legal issues that will arise from AI. (Cf. SPC, 2015; Ding, 2018)

Tencent researchers and CAICT academics concluded that a possible Chinese leadership in AIE and safety could be a way for China to accomplish the strategic competitive advantage or high ground. The stated that "China should actively construct the guidelines of AI ethics to play a leading role in promoting inclusive and beneficial development of AI. We should actively explore ways to go from being a follower to being a leader in areas such as AI legislation and regulation, education and personnel training, and responding to issues with AI." (TRI, 2017)

An important sign of ambitions by China to shape AI standards is the involvement with the International Organization for Standardization / International Electrotechnical Commission (ISO/IEC) Joint Technical Committee (JTC). It is one of the most prolific and largest technical committees regarding international standardization, which formed a special committee on AI in 2017 named ISO/IEC JTC 1/SC 42. They published 8 ISO standards and have 23 ISO standards under development. (Cf. ISO, 2021) The chair of the committee is Wael Diab, who was Vice President of AI, IoT

Standardization and Strategy at Huawei until 2019, and the committee's first meeting was held in April 2018 in Beijing, China (cf. Diab, 2021a). The first meeting and the chair position were both vehemently contested affairs that in the end went China's way (cf. Ding, 2018) Diab also chairs the AI track of the 22nd Global Standards Collaboration meeting (GSC-22). At the Industrial Internet Consortium (IIC). (Cf. Diab, 2021b).

AIE research is conducted mainly through state-sponsored projects and initiatives by individual scholars. The **Chinese Academy of Sciences (CAS)** and the **Chinese Academy of Social Sciences (CASS)** are China's two leading research institutes under the guidance of the **State Council**. One project at CAS which is led by the Institute of Automation studies matters like the relationship between AI and humans. They particularly focus on challenges related to determining accountability. The researchers at CAS also explore practical problems, e.g., ethics issues caused by the placement of robots into families. (Cf. NLPR, 2018; Shaohua, 2019)

Various renowned scholars are particularly influential in driving forward ethics research. Duan Weiwen, Director of the Research Center for Science, Technology and Society at CASS, is one of the most prominent thinkers on philosophical, ethical and social issues of Big Data and AI in China. Duan repeatedly underlines that innovation develops faster than ethics. Therefore, demands distinct and specific work to deal with ethical risks in specific technology application scenarios instead of just abstract recommendations. He also promotes public involvement and supervision in the matter of ethics. (Cf. Weiwen, 2019; Jia, 2020a) Duan is also affiliated with the bilateral China-UK Research Centre for AI Ethics and Governance (ChinUK), which was founded in 2019 (cf. Zeng, 2019).

Other researchers approach AIE based on the perception of traditional Chinese philosophy. **Zeng Yi**, a CAS affiliated researcher, led the creation of the **Harmonious Artificial Intelligence Principles**. Those principles are based on the concept of harmony in Chinese philosophy and emphasize harmony between humans and machines. It advocates for a positive cooperation between the two. This concept is also present in the **Beijing AI Principles**. Zeng also drives major applied ethics research efforts in areas like brain-inspired neural network architectures. (Cf. Zeng, 2018b; BAAI, 2019; BRC, 2021)

Guo Rui, a Professor at Renmin University, is another renowned scholar and government advisor. He focuses on the translation of abstract ethical guidelines into actionable governance systems. He examines the ethical risks of specific Al applications, ranging from content recommendation and precision marketing algorithms to smart courts and sex robots. Guo has advocated for companies to set up **ethics committees** to mitigate risks that come with new technologies. (Cf. Nana, 2019)

Academia from China increasingly participates in global conversations on AIE. This can be seen in part as a result of the government calling on academia for an expansion of the discourse power in the field. The **Berggruen China Center** is a prominent example that intertwines the states soft power with scholarly exchanges. It was established by the Berggruen Institute and the Peking University in 2018. The stated objective of the center is to engage Chinese thinkers to "examine, share and develop ideas to address global challenges. One of the centers main research subjects is AIE. (Cf. BRC, 2018)

In 2020, the Tsinghua University established the **Institute for AI International Governance (I-AIIG)** to shape the discourse about AIE, being an active participant for China in AI international governance, and by "contributing wisdom to human civilization" (cf. I-AIIG, 2020, 2021).

Xue Lan, Director of Tsinghua's said institute, has warned that geopolitical tensions between the US and China are having a troubling impact on policy exchanges and industry in the AI field. He reckons that this may obstruct beneficial collaboration on global AI governance. Lan compares AI to Pandora's Box. Stating that "it may become the last invention of mankind" in a scenario in which "it is not well controlled". (Cf. Lan, 2020)

The following illustration shows the number of peer-reviewed Ai publications in China by institutional affiliation from 2000 to 2019. (Zhang et al., 2021, p. 21)

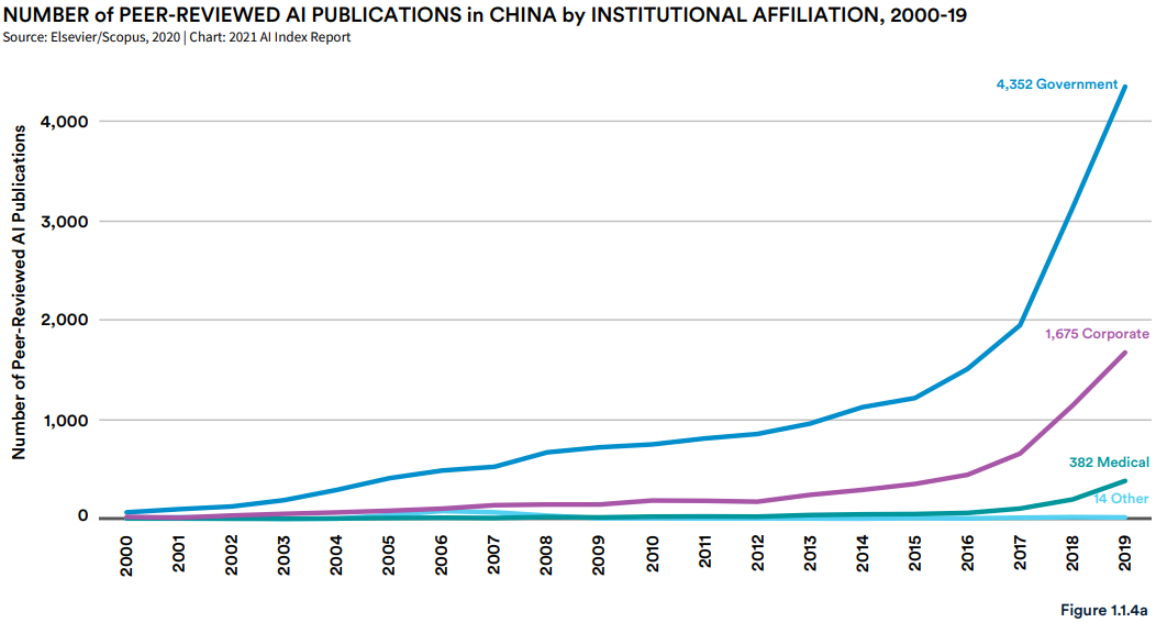

Figure 7 - AI publications in China by institutional affiliation from 2000 to 2019 (Zhang et al., 2021, p. 21)

4.2.4. Industry

The industry is seen as a vital force in organizing self-regulation, research, and education on AIE by China's leadership. Although state authorities are ultimately set the rules for governance. The state has stressed the significance of the industry to self-regulate to a certain extent. This was underlined with a paper published by the CAICT, which identified corporations as the main entities for AI governance short-term. (Cf. CAICT, 2020)

Chinese companies are involved in initiative regarding AIE at leading international standards associations (cf. Ding, 2018). Several prominent technology corporations and startups have announced initiatives to deal AI governance and ethics. They focus on concerns related to the development and commercialization of applications that include AI. While initiating their own research and introducing their own principles to tackle ethics issues, companies are also joining multistakeholder attempts to develop ethics principles and industry standards for **responsible AI**. Numerous companies were indirectly or directly engaged in each of the most relevant AI documents. The collaborative pledge of those documents was a commitment by the industry to **self-regulate**. Two executives from facial recognition unicorn Megvii and e-commerce behemoth JD.com are members of **MOST's** seven

member compromised AI governance committee. That demonstrates that the industry is directly participating in the creation of guiding documents and policy recommendations such as the **Governance Principles**. The formerly mentioned BAAI and other key industry alliances behind AI principles are funded by members from AI startups and Tech giants. (Cf. BAAI, 2019)

Tencent and Baidu are such giants which have also proposed AIE recommendations directly to state authorities (cf. Li, 2019; Jing, 2019). Self-regulation in the industry has predominantly led to high-level ethics codes. The before mentioned Tencent, Baidu, and Megvii have released documents that set forth ethics principles to guide their own and the general AI development. They equally underline the same principles such as data privacy, human oversight, accountability, and technical robustness and safety. The AI principles developed by Tencent are one of the most comprehensive in comparison in the Chinese industry. Published in 2018, they call for AI to be controllable, comprehensible, reliable, available, and they especially emphasize on algorithmic transparency. (Cf. Cao, 2020)

The Chinse industry directs extensive research into AI governance and ethics issues. Bigger companies to that with special distinct departments. Their work varies from researching methods for the preservation of privacy in ML to techniques for the protection against adversarial attacks on deep learning algorithms. A lot of that research preceded the increased attention of the government regarding AIE. (Cf. Ding, 2018)

Several AI executives and CEOs promote collaborative action and interdisciplinary exchanges on AIE. At key industry forums, e.g., Shanghai's annual World AI Conference some of those have positioned themselves as thought leaders on AI governance and ethics issues. Through initiatives like **AI for Good** they help boosting public awareness for AI risks in everyday life. (Cf. Tencent, 2019)

Many companies that have work in the AI space evidently acknowledge the significance of governing the societal and ethical impact of AI. Yet only some have established actions that turn high-level announcements or commitments into concrete procedures. There is no reason to suggest that their research and initiatives lack good faith intentions, but it can be noted that they often are deficient in the implementation of concrete measures that address the specific identified issues, e.g., data privacy or algorithmic fairness. (Cf. Arcesati, 2021)

One of the few companies that announced the creation of internal structures, i.e., an AIE Committee to oversee the implementation of its AI principles is the before mentioned **Megvii**. The committee has the objective to make recommendations to the board based on a whistleblowing procedure and internal investigations. It should be noted that one of the announced international members of the committee stated that he never joined. The effect of the committee remains so far unclear. (Cf. Megvii, 2019)

In 2018, **Tencent** released their ARCC principles, which stand for available, reliance, comprehensible, and controllable. They also released a report in 2019 on AIE in a digital society. (Cf. TAIL, 2019; Cao, 2020) The in 4.2.2. mentioned 2018 AI Standardization White Paper was co-authored, among others, by Alibaba, Tencent, Baidu, ByteDance, Huawei, and Intel China. (CESI, 2018)

Baidu was the only Chinese company entering the **Partnership On AI**. This international consortium consists of major players in the AI industry. Its objective is the development and establishment of best practices for AI systems for being socially beneficial. (Cf. Pai 2018) **Baidu** left the alliance in 2020 following tensions with the US. (Cf. Stone, 2020; Knight, 2020),

In 2018, **Baidu**'s CEO **Robin Li** faced intense resistance on social media when he argued that the Chinese population was more willing to trade it for convenience and less sensitive about privacy. The use of monitoring apps during the Covid-19 outbreak that collect location data and health information also prompted public criticism due to concerns over the erosion of privacy and discrimination. (Cf. Arcesati, 2021)

Consumer backlash has played a key role in holding Chinese tech companies **accountable** for **data privacy** violations and urging on regulators to create more strict regulations in recent years. The resulting data regime imposes far reaching restrictions on companies' ability to handle personal information. Restriction will be facilitated through the Personal Information Security Specification and the Personal Information Protection Law. (Cf. Lee et al., 2021)

4.3. EU

4.3.1. Introductory Considerations

The EU as a union of states is not represented in the AI Readiness Index. The 27 member states are individually ranked between 3 and 61 out of 172. The best ranked are Finland (3), Germany (4), and Sweden (5), and the worst ranked are Bulgaria (50), Croatia (58), and Greece (61). (Cf. Shearer et al., 2020, pp. 128-130)

There are around 5,172 AI startups in the EU based on data of each EU member state between April 2020 and September 2021. The most can be found in Germany (1,083) and France (941). Per capita most can be found in the Estonia (9.2 per 100.000 inhabitants). (Cf. Tracxn, 2021c; Simmonds, 2021; Kendall, 2021)

The EU as long ongoing ambitious projects regarding AI, i.e., the Human Brain and SPARC Projects. Already in 2013, they proposed a ten-year Human Brain Project, which is one of the most significant human brain research projects globally. (Cf. TRI, 2017)

Surveys of European perspectives regarding AI have shown that citizens hold a mostly positive view of developments with such technologies, viewing them as a positive addition to their lives, the economy, and overall society. Those attitudes highly vary by gender, age, location, and educational level, as well as largely dependent on someone’s exposure to AI and relevant information. (Cf. EC, 2012, 2017a)

Overall, a rather skeptical basic attitude toward algorithms was revealed. 36% of respondents saw more risks than opportunities in algorithmic decisions. 46% were undecided and 18% saw more opportunities than risks. 37% simultaneously considered them scary and incomprehensible and 35% feared a loss of control. Specifically, 57% respondents feared that programmers of algorithms have too much power over users, 68% feared that companies collect too much data about people, and 55% feared that algorithms can be easily manipulated. (Cf. Fischer & Petersen 2018, pp. 17-19)

In a Europe-wide representative population survey conducted in September 2018, i.e., 8 months after the survey by Fischer et al. in 2018 and differing slightly from the latter in methodology as well as in the wording of the questions, 40% of the Germans surveyed saw more benefits from algorithms and 24% saw more problems, while 36% were undecided. (Cf. Grzymek & Puntschuh, 2019, p. 24)

Only 22% would be happy to drive in a driverless car. Suspicious of social media, with only 7% viewing content on social media as generally trustworthy. Respondents were overall worried that digitization and automation would lead to job losses, and skeptical whether it would increase job opportunities across the EU. (STOA, 2020)

4.3.2. Government

In a 2017 Interim Report on Digital Single Market, the European Commission identified a need to assess whether there may be a need to adapt the current regulatory framework to new technological developments such as AI (cf. EC, 2017b, p. 14).

On April 25, 2018, the European Commission adopted its Communication on AI for Europe, which sets out the EU's AI initiative to exploit the opportunities of AI and minimize or prevent the emerging challenges of AI (cf. EC 2018). This AI initiative provides for the implementation of the actions, which are divided into three parts: (1) Promoting the EU's technological and industrial capabilities and the wider diffusion of AI throughout the economy through public and private investment; (2) Preparing for socio-economic changes; (3) Ensuring an appropriate ethical and legal framework. (Cf. EC 2018, pp. 7-22) The EU has the legislation in place to regulate AI application and address the challenges it faces. In May 2018, the first EU rules on network and information systems security, as well as stricter rules from DSGVO on personal data protection, were adopted. (Cf. EC 2018, p. 18)

The Commission established the **High-level Expert Group on Artificial Intelligence (AI HLEG)** in 2018. This group, which consists of 52 experts, acts as the steering group of the European AI Alliance. (Cf. EC, 2021b) The **European AI Alliance (EAAI)** is a forum engaged in a broad and open discussion of all aspects of Artificial Intelligence development and its impact (cf. EAAI, 2018).

In December 2018, AI HLEG published draft ethical guidelines for trustworthy AI. After further deliberations by AI HLEG in light of discussions on the European AI Alliance, stakeholder consultations, and meetings with Member State representatives, the guidelines were revised and republished in April 2019. (Cf. AI HLEG 2019) Regarding expert groups guidelines from the EU there is an identified increasing need to bring in legal and ethical expertise already at the programming stage. They also propose that more experts need to be trained. (Cf. AI HLEG 2019, pp. 23-24)

These guidelines do not have a binding character; therefore, they do not create legal obligations for Member States. Based on this document, the European Commission presented its own approach to the ethical aspects of AI application in its communication "Building trust in human-centric AI." (Cf. EC, 2019)

In their view, AI is "not an end in itself, but a tool that must serve people and ultimately enhance human well-being" (human-centered approach). Trust is a fundamental requirement for pursuing a human-centered approach to AI (trustworthiness of AI). (Cf. EC, 2019, p. 2)

The core principle of the EU Guidelines is that the EU must maintain a "human-centered" approach to AI that respects European values. The core requirements i.e., 7 Key Requirements for a Trustworthy AI for AI applications are defined as: (1) Primacy of human action and oversight; (2) Technical robustness and security; (3) Privacy and data quality management; (4) Transparency; (5) Diversity, non-discrimination, and fairness; (6) Social and environmental well-being; (7) Accountability. (Cf. EC, 2019, p. 4)

The European Parliament is also involved in AI legislation, adopting a resolution on comprehensive European industrial policy on AI and robotics in February 2019 (cf. EP, 2019). In this resolution, it called on the European Commission to periodically reassess existing legislation with a view to fostering a regulatory environment that is beneficial to the development of AI and consistent with the principle of better regulation, to ensure that it serves its purpose in relation to AI, while also respecting the EU's fundamental values, and to amend or replace new proposals where possible if this is demonstrably not the case. (Cf. EP, 2019)

In addition, this document also includes, among other things, in Section 4, the guidance or approaches to the creation of the future legal framework for AI, particularly with respect to the following aspects of AI application: (1) Personal data protection; (2) Liability issues; (3) Consumer protection and empowerment. (Cf. EP, 2019)

On February 19, 2020, White Paper "On Artificial Intelligence - A European Approach to Excellence and Trust" was published by the European Commission (cf. EC, 2020b). Stated in the White Paper among its goals to enable scientific breakthroughs, maintain EU technology leadership, and ensure that new technologies are at the service of all Europeans - bringing improvements to everyday life, while respecting citizens' rights (cf. EC, 2020b, p. 2). Furthermore, in this White Paper, the European Commission presented an approach for EU countries on how to harness the benefits of AI application for science, business, and society at large on the one hand, and how to overcome the challenges associated with it on the other. The White Paper also sets out the modalities of the future AI regulatory framework needed to create the "ecosystem for trust" in relation to AI use, as well as outlining how existing EU regulations could be adapted to take AI into account (fundamentally in product safety). (Cf. EC, 2020b, p. 3)

In the European Commission's view, the main issues to be considered are the risks to fundamental rights, safety, and the effective functioning of the liability regime (cf. EC, 2020b, p. 12). Along with the White Paper, European Commission published two companion documents on the same day, including a “Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics” and “A European strategy for data”. (EC, 2020a; 2020c)

On October 20, 2020, European Parliament adopted further decisions regarding a framework for the ethical aspects of AI (cf. EP, 2020).

The European Commission in April 2021 proposed regulation that establishes a regulatory structure focused on a risk-based classification of AI systems. This proposal also underlines the challenges of AI surveillance by noting that “[t]he use of artificial intelligence for the purposes of indiscriminate surveillance of natural persons should be prohibited when applied in a generalized manner to all persons without differentiation (…)”. (Cf. EC, 2021a)

In addition to EU led initiatives and regulation, there are also relevant developments inside many of the member states – often intertwined with or funded by the EU. One notable is AI Portugal 2030, a Portuguese national initiative on digitalization, an innovation and growth strategy to foster Artificial Intelligence in Portugal in the European context. (Cf. INCoDe.2030, 2021)

Manuel Heitor, Minister for Science, Technology and Higher education of Portugal, calls for AI that strengthens societal robustness. To achieve this objective, he argues that a clear vision of the impacts of AI on the labor market, democracy, fairness, security, privacy, equity, and commercial and governmental transparency must be established. The initiative identified ethics as "one of the most challenging aspects in AI", and advocates for AI that is made ethical-by-design to improve society and democracy. A specific action they call for is to have an ethics committee for AI that has participants from the public sector. (Cf. INCoDe.2030, 2021, p. 9, 13, 32)

4.3.3. NGOs & NPOs

The only partners from the EU in the before mentioned global **PAI** are the German Fraunhofer Institute for Industrial Engineering, and the Irish Insight Centre for Data Analytics from the University College Cork (cf. PAI, 2021).

In November 2016, the Green Digital Working Group of **The Greens/European Free Alliance**, published a position paper on Robotics and Artificial Intelligence. Related to AIE they emphasize on privacy as an inalienable human right and reject data ownership as a form of property right. They argue that if privacy is an inalienable human right, private data cannot be compromised or sold. They call for **liability** and **responsibility** but acknowledge that such approaches must be balanced to not put a heavy burden on AI start-ups and academia. (Cf. Albrecht et al., 2017)

The European Parliament launched **AI4People**, a multi-stakeholder forum with the objective to shape the social impact of AI applications, in 2018. It was originally established by the Atomium – European Institute for Science, Media and Democracy (EISMD) in 2017, which has partnered with players from the industry, e.g., Audi, Microsoft, Elsevier, Facebook, and with civil society organizations, e.g., European Association for Artificial Intelligence, and Ada-AI. AI4People's work was the main inspiration for the in section 4.3.2. mentioned 7 Key Requirements for a Trustworthy AI presented by the European Commission in 2019 (Cf. EISMD, 2017) A result of this initiative was the 2018 released Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations by AI4People. (Cf. Floridi 2018)

Another initiative is the AI4EU consortium, which was established in January 2019, with the objective to create a European Artificial Intelligence On-Demand Platform and Ecosystem. The AI4EU Platform wants to encourage discussion on the Ethical, Legal, Socio-Economic, and Cultural (ELSEC) aspects of AI. It's funded in part by the European Union's Horizon 2020 research and innovation program. (Cf. AI4EU, 2019)

Charlotte Stix, an experienced technology policy expert with a specialization in AI governance, recently concluded that "[t]he road towards a proportionate regulation for AI and implementing it within the ecosystem is still long and rocky, but the EU is taking steps towards achieving this goal." Eva Kaili, Chair of the (STOA) and the Centre for Artificial Intelligence (C4AI), argues that "Europe could lead as a global rules and standards setter for the Fifth Industrial Revolution." (Cf. Hui & Tse, 2021)

The **Stockholm International Peace Research Institute** (SIPRI) report on the impact of AI on strategic stability and nuclear risks warns against increasing use of autonomous or AI-based decision support systems that only appear to provide a clear picture in a short time. To maintain a degree of stability, they say, exchanges among militaries on respective AI capabilities are necessary to maintain the principle of nuclear deterrence. (Cf. Boulanin, 2019, p. 50 f.)

The Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE) wants to facilitate a European Vision for AI. CLAIRE aims to create a pan-European network of centers of excellence in AI strategically located throughout Europe. Related to AIE, CLAIRE wants to focus on trustworthy AI that enhances human intelligence rather than replacing it, thus benefiting the people of Europe. It was publicly launched with a letter of intent in 2018 and got financial support from the European Commission when it was incorporated in 2020. They contributed among other things by providing feedback to the white paper and proposal on AIE from the European Commission, talked about in section 4.3.2. The initiative has broad support from more than 1000 AI experts. In addition, 10 members of the AI HLEG are also CLAIRE supporters. (Cf. CLAIRE, 2018, 2020, 2021; Krempl 2018)

AlgorithmWatch is a nonprofit organization with the goal of looking at and classifying processes of algorithmic decision making that have societal relevance – that is, that either predict or predetermine human decisions, or make decisions in an automated fashion. They created a library for global AIE guidelines, which is comprised of 173 guidelines (last update in April 2020). (Cf. AlgorithmWatch 2021)

The following illustration shows the number of peer-reviewed Ai publications in the EU by institutional affiliation from 2000 to 2019. (Zhang et al., 2021, p. 22)

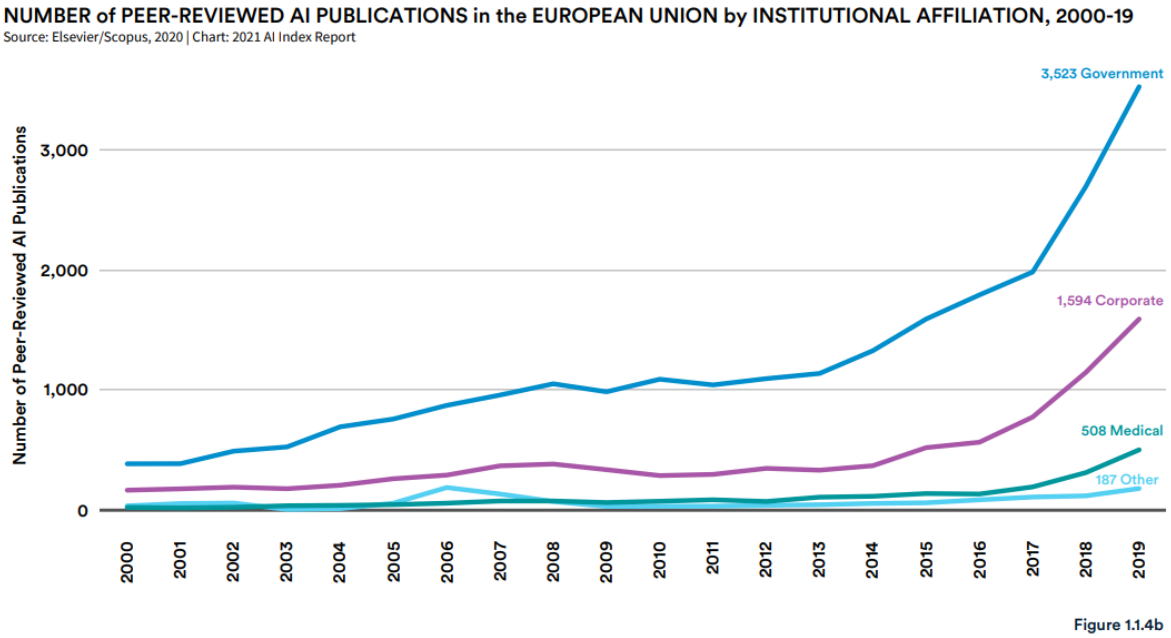

Figure 8 - AI publications in the EU by institutional affiliation from 2000 to 2019 (Zhang et al., 2021, p. 22)

4.3.4. Industry

SAP's Guiding Principles for Artificial Intelligence have been released by SAP in 2018. They defined seven principles, which are: (1) We are driven by our values, (2) We design for people, (3) We enable business beyond bias, (4) We strive for transparency and integrity in all that we do, (5) We uphold quality and safety standards, (6) We place data protection and privacy at our core, (7) We engage with the wider societal challenges of AI. The principles have not changed since then. Besides defining theoretical principles, SAP also established an AI Ethics Steering Committee and AI Ethics Advisory Panel. The principles were formulated by the AI Ethics Steering Committee. While the AIE Steering Committee consists of SAP executive personnel, the AIE Advisory Panel consists of academics, policy experts, and industrial experts that are not part of SAP such as Peter Dabrock, Chair of Systematic Theology of the University of Erlangen-Nuremberg, and Susan Liautaud, Lecturer in Public Policy and Law of Stanford University. SAP was the first European tech company that established an external Ethics Advisory Panel for AI (Cf. SAP, 2018a; SAP, 2018b; SAP 2021)

Markus Noga, VP ML at SAP, was appointed to the AI HLEG by the European Commission. Pekka Ala-Pietilä, SAP Independent Board Member, is the Chairmen of the AI HLEG. That way SAP directly linked to the AI HLEG and contributes to the EU's discourse on AI. Noga is now VP Solutions Technology at Suse but still a member of the AI HLEG. (Cf. EC, 2021b)

Luka Mucic, SAP CFO, said that "while the scope of the principles may be similar to what other companies have been doing, our commitments surrounding the principles set an important precedent for the industry that we believe can serve as a template for other companies" (cf. SAP, 2018b).

The 2020 established **Bosch code of ethics for AI** consists of five principles: (1) All Bosch AI products should reflect our "Invented for life" ethos, which combines a quest for innovation with a sense of social responsibility. (2) AI decisions that affect people should not be made without a human arbiter. Instead, AI should be a tool for people. (3) We want to develop safe, robust, and explainable AI products. (4) Trust is one of our company's fundamental values. We want to develop trustworthy AI products. (5) When developing AI products, we observe legal requirements and orient to ethical principles. Besides the formulation of principles, Bosch also defined approaches in the decision-making process with AI. In section 2.1.1. described Human in the Loop (HITL) is one of them. They Furthermore describe the Human in command (HIC) and the Human on the loop (HOTL). The AI system is used solely as a tool in the HIC approach. Meaning that personnel in charge always decides how and when the results presented by the AI tool shall be used. The HOTL approach is relevant for cases in which professionals during the design process have defined the relevant parameters for decisions, but the decisions themselves are assigned to the AI system. (Cf. Bosch, 2020) Christoph Peylo is the Global Head of Bosch Center for Artificial Intelligence (BCAI) and a member of the AI HLEG (cf. EC, 2021b).

Volkmar Denner, CEO Bosch, said in 2020 that "[w]e have to not only develop AI but build trust in AI as well." He argues that "it's time to stop being hysterical about artificial intelligence" and that "[w]e need to focus on its benefits in everyday life." (cf. Denner, 2020)

Tieto (since 2019 TietoEVRY) established AIE guidelines in 2018. Besides introducing guidelines, Tieto also introduced **internal ethics certification** to ensure that those working with AI systems adhere to the ethics guidelines. Furthermore, they committed to setting up new positions that focus on ethical values embedded in AI. Christian Guttmann, VP and Head of Artificial Intelligence and Data Science at Tieto, argued that "[b]y ingraining AI ethics principles and certification in our daily operations, we believe we can contribute to a more sustainable future, and ensure we build responsible AI that supports humanity." By his judgment, "[c]ompanies developing AI have an important role to play in shaping AI ethics and we need to listen to all sections of society to ensure the highest standards and benefits". Additionally, Tieto aims to create an ethical AI framework that should address AIE questions, and they contribute to AI initiatives, such as the **European AI Alliance** and the **CLAIRE** project, which both are described in section 4.3.3. (Cf. TietoEVRY, 2018)

In 2018, **Deutsche Telekom's** introduced its guidelines for artificial intelligence. They defined nine self-binding guidelines: (1) Responsible, (2) Careful, (3) Supporting, (4) Transparent, (5) Secure, (6) Reliable, (7) Trustworthy. (Cf. Deutsche Telekom, 2018)

Other fairly similar EU industry AIE principles or guidelines got released by **Telefónica** and **OP Financial Group** in 2018 (cf. Telefónica, 2018; OP, 2018)

A noteworthy technical solution is offered by **Brighter.ai**, a deep tech startup founded in Berlin, Germany, in 2017. They have developed a process to anonymize personal characteristics captured by cameras, such as faces. Their technology seeks to ensure compliance with privacy policies and regulatory requirements such as the GDPR while preserving camera data for ML approaches. (Cf. Brighter.ai, 2021)

5. CASE STUDY DISCUSSION

Collaboration between countries on regulations and standards will be essential to ensure a beneficial AI future. Given China's fast improvements in AI applications and the expanding global reach of its companies, it will be crucial for the EU and the US to engage with Chinese entities. Broad similarities between the ethical interests of all three markets could become the fundament for constructive collaboration. Several parts of MOST's Governance Principles strongly resemble EU or US guidelines, which indicates that Chinese researchers are already assessing EU and US recommendations and perceptions regarding AI. Resemblance between terms does not imply similarity in meaning. China's Governance Principles roughly resemble the EU's and US's Ethics Guidelines by emphasizing on environmental sustainability, conformity to human values, explainability, fairness and non-discrimination, technical safety and robustness, and respect for privacy.

Cross-cultural cooperation is essential for various reasons. It enables researchers to share resources, best practices, and expertise. This allows faster progress on managing the safety and ethical issues that may arise, and the development of beneficial AI applications. Cooperation will be essential in making sure that no part of society is disproportionately negatively impacted by AI. Without cooperation, the risk increases that competitive pressures between countries lead to AI developments that are less ethical, safe, and socially beneficial. (Cf. Askell et al., 2019; Ying, 2019; ÓhÉigeartaigh et al., 2020)

The need for international cooperation also arises through the applications of AI that are set to cross markets, e.g., autonomous vehicles. Such applications need to be able to interact well with a variety of different regulatory ecosystems and other technologies in several regions. (Cf. Cihon, 2019).

Rebecca Arcesati, an analyst at the Mercator Institute for China Studies (MERICS), argues that one must understand the terms that are used by the government to determine their vision for AI governance and ethics. Terms like "human rights" and "societal security" for instance do not imply the endorsement of liberal democratic values or individual freedoms. It rather must be seen in the context of maintaining stability by prioritizing collective wellbeing. Furthermore, she concludes that the aim for human-machine harmony alongside the call for boosting the guidance of public opinion guidance can be seen as an indication of the objective to make the society ready for larger governance and data-driven monitoring through AI. (Cf. Arcesati, 2021)

Yi Zeng, Director of CAS, argues that global cooperation on AI is not a Zero-Sum Game. He declares that "[t]o realize the global development of AI technology and its applications to serve the well-being of humankind and a better future for all of us, we must stand and hold together as a human community with a shared future." (Zeng, 2021)

Stuart Russell, a recognized American scholar in the field of AI, argues that the AI community has not yet adjusted to the fact that we are now starting to have a really big impact in the real world. Until now, AI development has primarily taken place in the lab, so the question of real-world impact hasn't been a concern. He argues that we now have to grow up very quickly to catch up. Russell's concerns relate to the so-called alignment problem of artificial intelligence: With increasing autonomy, AI systems must act as precisely as possible along with human needs. (Cf. Russell, 2021)

Rogier Creemer, Assistant Professor in Modern Chinese Studies, concludes that in China, ethical considerations regarding algorithmic decision-making are rather outlined around the interest of the collective instead of the individual (cf. Creemer 2017) There is an emerging data protection regime to impose restrictions on companies regarding the collection of personal information, but it leaves the government with “nearly unrestrained power to harvest and us citizens data for public security and law enforcement” (cf. Arcesati, 2021).

China intends to be a major force in global AI governance and ethics developments. The Governance Principles by the MOST calls for "a broad consensus on an international AI governance framework, standards and norms" and advocates for borderless “open collaboration”. China's objective to strengthen international research and to push a consensus regarding shared AI challenges are expressed by the AIDP. (Arcestasi, 2021) The involvement of China in intergovernmental AI governance endeavors is limited. Several multilateral attempts lack Chinese contribution. This might be due to the given emphasis on human rights and democratic values on in endeavors like the **OECD Principles on AI** and **Global Partnership on AI**. (OECD, 2019; 2020) The participation of Chinese actors in international initiatives is increasing. China, as a member of the G20, signed the group's non-binding AI principles which are based on the OECD principles (cf. G20, 2019) Chinese experts directed a consensus at the UNESCO on AI for education and engaged in an expert group which drafted recommendations for AIE (cf. UN, 2020).

On issues like the use of facial recognition during the Covid-19 pandemic, Chinese scholars carried out research with international counterparts (cf. Zeng, 2020) Researchers from the AI industry engage in global research projects on technical challenges (cf. Cao et al., 2019). For the EU, “human-centric” is a cornerstone of their approach to AIE. (AI HLEG, 2019; EC, 2019, 2020) EU principles are being analyzed by Chinese academics in the context of specific applications, e.g., education and medical (cf. Shen & Wang, 2020; Xiang et al. 2020).

China is not the only one using or exporting AI for surveillance, involving practices that raise severe ethical concerns (cf. Feldstein, 2019) The scale of the nation's ambitions to utilize AI to strengthen an authoritarian governance system is what sets China apart. One can make the case that the EU and the US should prioritize working with like-minded democracies to develop standards that are embedded in liberal democratic values.

There is no doubt that the, e.g., in section 4.1.4. described unrightful usage of private medical information is generally wrong. Still, it also seems evident that the use could have potentially saved or helped many lives. The measurements by SAP are something more companies should adopt. Some other companies mentioned in this paper might have established similar measures. Having internal and external supervision in addition to internal ethics certification, and the contribution in governmental and non-governmental entities should be the baseline for large tech enterprises that are involved with AI.

6. CONCLUSIONS

6.1. SYNTHESIS OF THE DEVELOPED WORK

The main objective of this thesis was to analyze the landscape of AIE. Based on the findings in the literature review, different markets have been analyzed and compared focused on specific factors. The core synthesis of this thesis are the results of the Case Study, which have been laid out in the previous section. Therefore, this synthesis will mainly consist of a short reflection on the Goals and Questioned defined in section 1.2., only Question 4 will be emphasized more detailed, because it has not been so far.

Goal 1 - Identifying and analyzing challenges in AI Ethics, was reached through section 2.1. Challenges, among others, are AI-based discrimination, traceability, explainability, and black-box. Cross-cultural cooperation could be the key to those challenges but based on the current political climate between the US and China, it does not seem realistic to expect any significant collaboration soon.

Goal 2 - Identifying harm and discrimination through the lack of AI Ethics, was covered in section 2.1.3. to 2.1.7. Harms are allocational or representational, and discriminations are direct, indirect, or intersectional.

Goal 3 - Identifying and comparing AI Ethics in different markets, was reached after section 4. and 5.

Goal 4 - Raising questions for further research, was taken care of in section 2.1.10 and 6.3., e.g., who decides on the development, production, and use of programs? And how does one deal with malfunctions or inhumane consequences?

Question 1: What are the developments in the field of AI Ethics? Answered in section 2.1., e.g., frameworks, guidelines, initiatives, research, and conferences.

Question 2: What are the challenges in the field of AI Ethics? Answered with Goal 1.

Question 3: How do different markets adjust to AI Ethics? Answered in section 4.

Question 4: What is the view of industry leaders on AI Ethics? Answered in section 4.1.4., 4.2.4., and 4.3.4. by presenting a variety of viewpoints from leading companies and relevant individual. Being on the front of discovering and addressing the harmful impacts of AI applications is obvious for the AI industry because they are the ones funding research, as well as developing and deploying AI in real-life conditions. As the providers of AI products and services, they are incentivized to address the risks to prevent backlash from the public and regulators. Whether AI ethics declarations from the industry are leading to meaningful changes in the research and development processes, or whether they are rather empty commitments that serve mainly as a means to increase the reputation of a company is still unclear. Usually, companies are hesitant to implement time-intensive and costly procedures to ensure ethical AI products. An additional layer of complication and a reason for skepticism is the close relationship between the AI industry and the government. In this relationship, the government provides the policy but is quite often one of the major customers of the AI industry that they are trying to regulate. The pledges on AIE by companies often stand in severe contrast to the AI products and services that they sell, e.g., facial recognition or tools for analytics to the public security system.

6.2. LIMITATIONS

One of the biggest limitations of this work is the amount of new research and guidelines nearly every day. The state-of-the-art aspiration of this paper was formulated at the beginning of 2021. The author tried to keep up with new developments in academia, governments, and the industry, but it should be noted that new developments are fast-paced and manifold. This is primarily a limitation regarding the Case Study comparison. The problem that arises with this limitation is that information collected regarding one market might be more up to date than another.

Another limitation is the language barrier in regard to the research regarding China. The strong dependence on Google translate or secondary source materials has put a limitation on the state-of-the-art aim of this research. Both those limitation get even worse when combined because the newest research and developments in China are mainly first only published in Chinese. There is also the risk of misinterpretation when algorithmic translation or secondary sources are used.

The broad scope of the paper might have created a scenario in which many topics have been touched, but not many have been analyzed deeply and thoroughly. A smaller scope might have been a better option to gain significant scientific value.

Market limitations, e.g., Japan, Canada, and the UK are all relevant and acknowledged players in the AIE sphere but have not been talked about – only the UK regarding the Chin-UK initiative. This research is also limited or influenced by the author's own bias, e.g., confirmation bias and cultural imprint. The author has a bias in favor of liberal western democracies and has not lived or been in China.

Furthermore, the evaluation of all the principles, guidelines and frameworks described in this thesis is mainly on a surface level. Hence, the author relied heavily on the perspectives from expert in that field, which have their own agendas and bias. Government and companies can quite easily make over-inflated promises.

6.3. FUTURE WORK

For future work, the research should focus on specific sectors for AIE, e.g., healthcare, autonomous driving, surveillance, defense, misinformation, and deepfakes. A question could be whether deepfakes should better be banned entirely or if there are enough valuable and harmless use cases for them? Also, is it even feasible to ban such a technology? A particular focus can also be put on questions regarding liability and sustainability.

The work should be continued by comparing more markets. Based on the limitations, it is recommended to have teams assessing the markets in which there is at least one member that is native from the country and speaks the language.

The principles, guidelines and frameworks, and other promises by governments and the industry must be evaluated. It must be shown that the promises are not just empty phrases for reputation improvement. Many topics have been raised by this thesis of which many have the significance to be explored much further in detail. The author would especially like to further explore AIE in the military and evaluate the outcomes and acceptance of internal and external AIE Steering Committees or Advisory Boards in companies.

6.4. THE AUTHOR´S PERSPECTIVE

We are the first generation of people that give the power over decisions in the hand of machines. If we get it wrong, every generation that follows will pay the price for our mistakes. The challenges for AIE are similar to those for human ethics. The only difference is the scale of the possible good or harm. Some of the most pressing issues surrounding AI are already being addressed.

The issue of new weapons systems cannot be solved by technology alone, but it is an important issue that will need to be addressed. The development of autonomous weapons will likely follow the development of autonomous vehicles. AI is already used in warfare, e.g., in the military, to support the targeting of weapons. The solution must be political and must be addressed in international law. Surveillance to a certain degree is necessary for security, but the technology could be misused or abused. The more powerful the technology, the bigger the risk of being abused or in the wrong hands. There is a fine line between surveillance and security. In the future, the challenge will be to keep this line from being crossed. Differences between generally prioritizing the individual or the collective will impact the developments in the different markets.

The real challenge for the next 20 years will be the integration of AI into everyday life. To that end, we need to have a deeper understanding of the capabilities and limits of AI. This understanding could also help us develop more robust ethical frameworks. By understanding the nature of these challenges, we can take steps to avoid them. Through this deeper understanding, we could be enabled to make better ethical choices. The potential benefits of AI are too great to be stifled by excessive regulation, but that does not mean that AI should be unregulated. AI should be regulated in terms of privacy and security. Governments should regulate the use of AI in surveillance and security, cybercrime, and algorithmic bias.

AI Startups need freedom, but also guidance. A comprehensive AIE framework could give that guidance. The more countries adapt a framework collaboratively, the better for AI startups.

Government should focus on promoting AI and educating about its opportunities while being transparent about the challenges. This could be best done by funding AI research. Funding for research in AI should increase as it will help us understand how AI works. Exploring the potential dangers of AI as part of that research is essential to support AI safety and enable acceptance for AI, which might lead to an overall positive public sentiment. Educating the public about AI is critical for success. The public will need to understand the benefits and risks of AI. If there is no fundamental understanding, AI can be easily blamed for complex issues or inconvenient outcomes.

Ethical AI will only arrive with true emotional awareness in machines. Meaning that to understand the ethical implications of their actions truly, machines will need to feel and empathize. That is something that might be hard for many people to accept, but the more we learn about the nature of consciousness and the role of emotions and feelings, the more we are forced to take this as a genuine possibility. Until that happens, the ethical AI will remain in the hands of humans.

I would argue that it is often better and necessary to take a step back when alleged bias is discovered to explore the causes and results thoughtfully. We live in a time in which public outrage is a daily business. Enabled by social networks, headlines get treated as facts, and participants on various sides of the political spectrum are keen to act fast, but often in bad faith and driven by dogmatic ideologies. With the debate about AI and its harms shifting more into the mainstream, it can be expected that with it will come an activist and ideology-driven dispute about whether equity should be a core principle of AIE or not. We are more sensitive to bias and discrimination – especially when it comes to sex and ethnicity than ever before. This is, of course, an astonishing development, but it can come with irrational fears when it meets technology.

Using AI wrong is problematic, but not using AI is also problematic in many areas, because not using it at all – missing out on possibilities that come with it – has a more significant downside than the harm that come with the use. We should accept, to a certain degree, that we will find AI bias and discrimination. We should be careful in framing those findings as if the bias results from AI itself or is more significant than human bias. Nonetheless, we should try to mitigate all forms of bias and discrimination – be it human or artificial.

After researching the subject, the author has more questions than before. The two most pressing are: Do we want AI to represent reality as precisely as possible, or do we want AI to represent a reality that we want to decrease a feedback loop of bias? Is, e.g., the google search engine discriminatory when it shows us the results that it 'thinks' are the ones most fitting to our query without considering that the result might enhance a bias?

There is no truth API; there is no ethics API. In the end, truth and ethics are in the hands of power, and human rights issues are intertwined with political interests and geopolitical strategies. The greatest challenge for Artificial Intelligence is Natural Stupidity. Besides nuclear power and climate change, AI might be the topic for which global collaboration is needed the most to navigate humanity safely through the 21st century and beyond.

I want to thank my supervisor, Prof. Dr. Vitor Duarte dos Santos, for guiding me through the adventure of writing this thesis, and I would like to use my last words to thank Rita Gonçalves for being the most significant source of joy and love during those sometimes exhausting and bizarre months.

BIBLIOGRAPHY

Accenture (2018). Accenture Launches New Artificial Intelligence Testing Services. Retrieved from https://newsroom.accenture.com/news/accenture-launches-new-artificial-intelligence-testing-services.htm

AI HLEG (2019). Artificial Intelligence High-Level Expert Group. Ethics Guidelines for Trustworthy Al. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

AI4EU (2019). Artificial Intelligence for the European Union. About AI4EU. Retrieved from https://www.ai4europe.eu/about-ai4eu

AITO (2021). Artificial Intelligence & Technology Office. Cybersecurity Awareness: Adversarial AI Attacks. Retrieved from https://www.energy.gov/sites/default/files/2021-10/Cyber%20Awareness%20-%20Adversarial%20AI%20v3.pdf

Albrecht, J. P., Reda, J., Andersson, M., Reimon, M. & Reintke, T. (2017). Position on Robotics and Artificial Intelligence. Retrieved from https://juliareda.eu/wp-content/uploads/2017/02/Green-Digital-Working-Group-Position-on-Robotics-and-Artificial-Intelligence-2016-11-22.pdf

AlgorithmWatch (2021) AlgorithmWatch. Retrieved from https://algorithmwatch.org/

Anderson, K., & Waxman, M. C. (2013). Law and ethics for autonomous weapon systems: Why a ban won't work and how the laws of war can.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. AI magazine, 28(4), 15-15.

Anderson, M., & Anderson, S. L. (Eds.). (2011). Machine ethics. Cambridge University Press.

Anderson, M., Anderson, S., & Armen, C. (2005, November). Towards machine ethics: Implementing two action-based ethical theories. In Proceedings of the AAAI 2005 fall symposium on machine ethics (pp. 1-7).

Anderson, S. L., & Anderson, M. (2015). Towards a principle-based healthcare agent. In Machine Medical Ethics (pp. 67-77). Springer, Cham.

Angwin, J. et al. (2016a). Machine Bias. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Angwin, J. et al. (2016b) How We Analyzed the COMPAS Recidivism Algorithm. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Arcesati, R. (2021). Lofty Principles, Conflicting Incentives. Mercator.

Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research [research frontier]. IEEE computational intelligence magazine, 5(4), 13-18.

Askell, A., Brundage, M., & Hadfield, G. (2019). The role of cooperation in responsible AI development. arXiv preprint arXiv:1907.04534.

Audi, R. (1999). The Cambridge dictionary of philosophy.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. Nature, 563(7729), 59-64.

Awad, M., & Khanna, R. (2015). Efficient learning machines: theories, concepts, and applications for engineers and system designers (p. 268). Springer nature.

BAAI (2019). Beijing Academy of Artificial Intelligence. Beijing AI Principles. Retrieved from https://www-pre.baai.ac.cn/news/beijing-ai-principles-en.html

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671.

Bartlett, J. (2018). The People vs Tech: How the Internet is killing democracy (and how we save it). Random House.

Basu, A. (2018). Discrimination in the Age of Artificial Intelligence.

Batra, G., Jacobson, Z., Madhav, S., Queirolo, A., & Santhanam, N. (2019). Artificial-intelligence hardware: New opportunities for semiconductor companies. McKinsey and Company, January, 2.

Baumann, F., Lorenz-Spreen, P., Sokolov, I. M., & Starnini, M. (2020). Modeling echo chambers and polarization dynamics in social networks. Physical Review Letters, 124(4), 048301.

Baxter, P., & Jack, S. (2008). The qualitative report qualitative case study methodology: Study design and implementation for novice researchers (Vol. 13).

Beck, S. (2020). Künstliche Intelligenz–ethische und rechtliche Herausforderungen. Philosophisches Handbuch Künstliche Intelligenz, 1-28.

Beck, U. (1988). Gegengifte. Frankfurt: Suhrkamp.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2020). On the linguistic representational power of neural machine translation models. Computational Linguistics, 46(1), 1-52.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.

Bendel, O. (2019). The synthetization of human voices. Ai & Society, 34(1), 83-89.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604.

Bentham, J. (1879). The principles of morals and legislation. Clarendon Press. Original work published in 1789.

Bigelow, K. (2018) Ethics in healthcare aren't new, but their application has never been more important. Retrieved from https://www.gehealthcare.com/article/ethics-in-healthcare-arent-new-but-their-application-has-never-been-more-important

Binns, R. (2018a). Fairness in machine learning: Lessons from political philosophy. In Conference on Fairness, Accountability and Transparency (pp. 149-159). PMLR.

Binns, R., van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018b). 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 Chi conference on human factors in computing systems (pp. 1-14).

Blodgett, S. L. (2021). Sociolinguistically Driven Approaches for Just Natural Language Processing.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. arXiv preprint arXiv:2005.14050.

Boddington, P. (2017). Towards a code of ethics for artificial intelligence (pp. 27-37). Cham: Springer.

Bogosian, K. (2017). Implementation of moral uncertainty in intelligent machines. Minds and Machines, 27(4), 591-608.

Bonifacic, I. (2021). Google pursues Pentagon cloud contract in spite of past employee concerns. Retrieved from https://www.engadget.com/google-jwcc-contract-214046745.html

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. Ethics and Information Technology, 20(1), 41-58.

Bosch (2020). In brief: Bosch code of ethics for AI. Retrieved from https://assets.bosch.com/media/en/global/stories/ai_codex/bosch-code-of-ethics-for-ai.pdf

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Botha (2019). The 15 most important AI companies in the world. Retrieved from https://towardsdatascience.com/the-15-most-important-ai-companies-in-the-world-79567c594a11

Boulanin, V. (2019). The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Euro-Atlantic Perspectives.

BRC (2018) Berggruen Research Center. The Third Understanding China Conference Co-Hosted by the Berggruen Institute 21st Century Council. Retrieved from https://www.berggruen.org/events/the-third-understanding-china-conference/

BRC (2021). Berggruen Research Center. Ethics in Digital Governance: Forum on the Ethics of Artificial Intelligence from a Global Perspective. Retrieved from https://www.berggruen.org/activity/ethics-in-digital-governance-forum-on-the-ethics-of-artificial-intelligence-from-a-global-perspective/

Brighter.ai (2021) About brighter.ai. Retrieved from https://brighter.ai/

Broman, M., (2017). Human Robotics/AI Interaction. Retrieved from https://www.katinamichael.com/istas17/2017/8/19/human-robotics-ai-interaction-by-morgan-broman

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

Bubinger, H., & Dinneen, J. D. (2021). Actionable Approaches to Promote Ethical AI in Libraries. Proceedings of the Association for Information Science and Technology, 58(1), 682-684.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Campolo, A., Sanfilippo, M. R., Whittaker, M., & Crawford, K. (2017). AI Now 2017 report.

Cao, J. (2020). "ARCC": An Ethical Framework for Artificial Intelligence. Retrieved from https://www.tisi.org/13747

Cao, Y., Xiao, C., Yang, D., Fang, J., Yang, R., Liu, M., & Li, B. (2019). Adversarial objects against lidar-based autonomous driving systems. arXiv preprint arXiv:1907.05418.

Carbon Calculator (2021) Carbonfootprint calculator. Retrieved from https://www.carbonfootprint.com/calculator.aspx

Carter (2021). China population: census confirms increase to 1.412 billion in 2020, but births fall again. Retrieved from https://www.scmp.com/economy/china-economy/article/3132980/china-population-latest-census-confirms-increase-1412-billion

Castelvecchi, D. (2016). Can we open the black box of AI?. Nature News, 538(7623), 20-23.

Cave, S., & ÓhÉigeartaigh, S. S. (2018, December). An AI race for strategic advantage: rhetoric and risks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 36-40).

CESI (2018) China Electronics Standardization Institute. Artificial Intelligence Standardization White Paper. Retrieved from https://www.aspi.org.au/report/mapping-chinas-technology-giants-reining-chinas-technology-giants

CESI (2018) translated with google. Retrieved from https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-key-chinese-think-tanks-ai-security-white-paper-excerpts/

Chaillan, N. (2021). US has already lost AI fight to China, says ex-Pentagon software chief. Retrieved from https://www.ft.com/content/f939db9a-40af-4bd1-b67d-10492535f8e0. (https://archive.is/dITl1)

China Daily (2019). Governance Principles for the New Generation Artificial Intelligence--Developing Responsible Artificial Intelligence. Retrieved from http://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html

Chou, J., Murillo, O. & Ibars, R. (2017). What the Kids' Game "Telephone" Taught Microsoft about Biased AI. Retrieved from https://www.fastcompany.com/90146078/what-the-kids-game-telephone-taught-microsoft-about-biased-ai.

Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89.

Churchill, L. R. (1999). Are we professionals? A critical look at the social role of bioethicists. Daedalus, 128(4), 253-274.

CIF (2020). China Innovation Fund. Publication of the "New Generation AI Governance Principles – Developing Responsible AI". Retrieved from http://chinainnovationfunding.eu/dt_testimonials/publication-of-the-new-generation-ai-governance-principles-developing-responsible-ai/

Cihon, P. (2019). Standards for AI governance: international standards to enable global coordination in AI research & development. Future of Humanity Institute. University of Oxford.

CLAIRE (2018). A European Vision for AI. Retrieved from https://claire-ai.org/vision/

CLAIRE (2020). Response to the European Commission White Paper - On Artificial Intelligence - A European approach to excellence and trust. Retrieved from https://claire-ai.org/wp-content/uploads/2020/06/ec-wp-response.pdf

CLAIRE (2021). Response to the European Commission's Proposal for AI Regulation and 2021 Coordinated Plan on AI. Retrieved from https://claire-ai.org/wp-content/uploads/2021/08/CLAIRE-EC-AI-Regulation-Feedback.pdf

Clarivate (2021). Clarivate's Web of Science. Retrieved from https://www.webofscience.com/wos/woscc/basic-search

Cox, I. J., & Wilfong, G. T. (2012). Autonomous robot vehicles.

Crawford, K. (2016). Artificial intelligence's white guy problem. The New York Times, 25(06).

Crawford, K. (2017, December). The trouble with bias. In Conference on Neural Information Processing Systems, invited speaker.

Creemer, R. (2017). Interview with Dr. Rogier Creemers: AI + Social Credit + Algorithmic Governance + Cybersecurity + VPNs. Retrieved from https://www.digitalasiahub.org/2017/08/14/interview-with-dr-rogier-creemers-ai-social-credit-algorithmic-governance-cybersecurity-vpns-cross-border-dataflows/

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. u. Chi. Legal f., 139.

Creswell, J. W., & Poth, C. N. (2016). Qualitative inquiry and research design: Choosing among five approaches. Sage publications.

Dalenberg, D. J. (2018). Preventing discrimination in the automated targeting of job advertisements. Computer law & security review, 34(3), 615-627.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 11, No. 1).

Davies, M., Srinivasa, N., Lin, T. H., Chinya, G., Cao, Y., Choday, S. H., ... & Wang, H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. Ieee Micro, 38(1), 82-99.

Davies, P. (2017) In: Brockman, J. Was sollen wir von Künstlicher Intelligenz halten?: Die führenden Wissenschaftler unserer Zeit über intelligente Maschinen. S. Fischer Verlag.

Daws, R. (2021). DeepMind hit with class-action lawsuit over NHS health data scandal. Retrieved from https://artificialintelligence-news.com/2021/10/01/deepmind-class-action-lawsuit-nhs-health-data-scandal/

Dencik, L., Hintz, A., Redden, J., & Treré, E. (2019). Exploring data justice: Conceptions, applications and directions.

Denner, V. (2020). Dr. Volkmar Denner, Bosch CEO, on the ethics of artificial intelligence. Retrieved from https://www.bosch.com/stories/denners-view-artificial-intelligence-ethics/

Derrida, J. (1967). Of Grammatology, corrected edition, trans. Gayatri Chakravorty Spivak (Baltimore and London: Johns Hopkins University Press, 1998), 84.

Deutsche Telekom (2018). Guidelines for Artificial Intelligence. Retrieved from https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366

Diab, W. (2021a) ResearchGate Profile. Retrieved from https://www.researchgate.net/profile/Wael-Diab

Diab, W. (2021b) Chair - ISO/IEC JTC 1/SC 42 Artificial intelligence. Retrieved from https://oecd.ai/en/wonk/contributors/diab-wael

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-dimensional gender bias classification. arXiv preprint arXiv:2005.00614.

Ding, J. (2018). Deciphering China's AI dream. Future of Humanity Institute Technical Report. & Chinese Interests Take a Big Seat at the AI Governance Table. Retrieved from https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinese-interests-take-big-seat-ai-governance-table/

DISER (2021). Department of Industry, Science, Energy and Resources. Australia's Artificial Intelligence Ethics Framework. AI ethics case study: Microsoft. Retrieved from https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework/testing-the-ai-ethics-principles/ai-ethics-case-study-microsoft

DoD (2019). U.S. Department of Defense. AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense

DoD (2020). U.S. Department of Defense. DOD Adopts Ethical Principles for Artificial Intelligence. Retrieved from https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

Doyle, O. (2007). Direct discrimination, indirect discrimination and autonomy. Oxford Journal of Legal Studies, 27(3), 537-553.

Dutton, T. (2018). An overview of national AI strategies. Medium. Politics+ AI. June, 28.

E.O. (2019). Executive Order 13859. Maintaining American Leadership in Artificial Intelligence. Retrieved from https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence

E.O. (2020). Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. Retrieved from https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/

EAAI (2018) The European AI Alliance. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/european-ai-alliance

EC (2012). European Commission. Special Eurobarometer 382: Public Attitudes towards Robots. Eurobarometer Surveys. Retrieved from https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/SPECIAL/surveyKy/1044/p/3

EC (2017a). European Commission. Special Eurobarometer 460: Attitudes towards the impact of digitisation and automation on daily life. Retrieved from https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/SPECIAL/surveyKy/2160

EC (2017b). European Commission. Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions on the Mid-Term Review on the implementation of the Digital Single Market Strategy. A Connected

Digital Single Market for All. Retrieved from https://eosc-portal.eu/sites/default/files/COM-2017-228-F1-EN-MAIN-PART-1.PDF. p. 14.

EC (2018). European Commission. Artificial Intelligence for Europe. COM (2018) 237. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

EC (2019). European Commission. Building Trust in Human-Centric Artificial Intelligence. COM(2019)168. Retrieved from https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

EC (2020a). European Commission. Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics. COM(2020) 64. Retrieved from https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1593079180383&uri=CELEX%3A52020DC0064

EC (2020b). European Commission. White Paper on Artificial Intelligence: A European Approach to Excellence and Trust. COM(2020)65. Retrieved from https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust

EC (2020c). European Commission. A European strategy for data. COM (2020) 66. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

EC (2021a). European Commission. Proposal for A Regulation Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). COM/2021/206. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

EC (2021b). European Commission. AI HLEG - steering group of the European AI Alliance. Retrieved from https://ec.europa.eu/futurium/en/european-ai-alliance/ai-hleg-steering-group-european-ai-alliance.html

Eckersley, P. (2018). Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). arXiv preprint arXiv:1901.00064.

Eggers, W. D., Schatsky, D., & Viechnicki, P. (2017). AI-augmented government. Using cognitive technologies to redesign public sector work. Deloitte Center for Government Insights, 1-24.

EISMD (2017). Atomium – European Institute for Science, Media and Democracy. AI4People – The first multi-stakeholder forum bringing together all actors interested in shaping the social impact of new applications of AI. Retrieved from https://www.eismd.eu/ai4people/

Engelmann, S., Chen, M., Fischer, F., Kao, C. Y., & Grossklags, J. (2019, January). Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines" Good" and" Bad" Behavior. In Proceedings of the conference on fairness, accountability, and transparency (pp. 69-78).

EP (2019). European Parliament resolution of 12 February 2019 on a comprehensive European industrial policy on artificial intelligence and robotics (2018/2088(INI)). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52019IP0081

EP (2020). European Parliament. Framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)). Retrieved from https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2020/2012(INL).

Ernest, N., Carroll, D., Schumacher, C., Clark, M., Cohen, K., & Lee, G. (2016). Genetic fuzzy based artificial intelligence for unmanned combat aerial vehicle control in simulated air combat missions. Journal of Defense Management, 6(1), 2167-0374.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. nature, 542(7639), 115-118.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. The Journal of Ethics, 21(4), 403-418.

Feigenbaum, E. A. & Feldman, J. (1963) Computers and Thought. American Association for Artificial Intelligence.

Feigenbaum, E. A. (2003). Some challenges and grand challenges for computational intelligence. Journal of the ACM (JACM), 50(1), 32-40.

Feldstein, S. (2019). The Global Expansion of AI Surveillance. Carnegie Endowment for International Peace. Retrieved from https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847.

Feng, C. (2021). Increasing use of facial recognition technology in China faces backlash from city governments. Retrieved from https://www.scmp.com/tech/tech-trends/article/3131442/increasing-use-facial-recognition-technology-china-faces-backlash

Fischer, S., & Petersen, T. (2018). Was Deutschland über Algorithmen weiß und denkt. Impuls Algorithmenethik.

Fjeld, J., Hilligoss, H., Achten, N., Daniel, M. L., Feldman, J., & Kagay, S. (2019). Principled artificial intelligence: A map of ethical and rights-based approaches. Berkman Klein Center for Internet & Society at Harvard University, 1.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. Public opinion quarterly, 80(S1), 298-320.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3), 330-347.

Fryer-Biggs (2018). The Pentagon plans to spend $2 billion to put more artificial intelligence into its weaponry. Retrieved from https://www.theverge.com/2018/9/8/17833160/pentagon-darpa-artificial-intelligence-ai-investment

G20 (2019). G20 Japan AI Principles. Retrieved from https://www.g20-insights.org/wp-content/uploads/2019/07/G20-Japan-AI-Principles.pdf

Gal, D. (2020). The AI powered state: China's approach to public sector innovation. China's Approach to AI Ethics. 53-61

Gardner, H. E. (2000). Intelligence reframed: Multiple intelligences for the 21st century. Hachette UK.

Garvie, C. (2016). The perpetual line-up: Unregulated police face recognition in America. Georgetown Law, Center on Privacy & Technology.

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665-673.

Geisler, N. L. (1989). Christian ethics (pp. 216-217). Baker Book House.

George, A. L., & Bennett, A. (2005). Case studies and theory development in the social sciences. MIT Press.

Gerring, J. (2016). Case study research: Principles and practices. Cambridge University Press.

Gershgorn (2015). New 'OpenAI' Artificial Intelligence Group Formed By Elon Musk, Peter Thiel, And More. Retrieved from https://www.popsci.com/new-openai-artificial-intelligence-group-formed-by-elon-musk-peter-thiel-and-more/

Gielscher, S., pies, I., Valentinov, V., & Chatalova, L. (2016). Rationalizing the GMO debate: the ordonomic approach to addressing agricultural myths. International journal of environmental research and public health, 13(5), 476.

Gillham, B. (2000). Case study research methods. Bloomsbury Publishing.

Glymour, B., & Herington, J. (2019, January). Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In Proceedings of the conference on fairness, accountability, and transparency (pp. 269-278).

Google (2004). What Would 2004 Google Say About Antitrust Probe?. Retrieved from https://www.wsj.com/articles/BL-DLB-33777

Google (2014). Alphabet replaces Google's 'Don't be evil' with 'Do the right thing'. Retrieved from https://www.engadget.com/2015-10-02-alphabet-do-the-right-thing.html

Google (2018). Artificial Intelligence at Google: Our Principles. Retrieved from https://ai.google/principles/

Google (2019). At I/O '19: Building a more helpful Google for everyone. Retrieved from https://www.blog.google/technology/developers/io19-helpful-google-everyone/

Greene, D., Hoffmann, A. L., & Stark, L. (2019, January). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In Proceedings of the 52nd Hawaii international conference on system sciences.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105-2108.

Greene, T. (2019). Report: Palantir took over Project Maven, the military AI program too unethical for Google. Retrieved from https://thenextweb.com/news/report-palantir-took-over-project-maven-the-military-ai-program-too-unethical-for-google

Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018b). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In Proceedings of the 2018 World Wide Web Conference (pp. 903-912).

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018a). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

Grzymek, V., & Puntschuh, M. (2019). Was Europa über Algorithmen weiß und denkt: Ergebnisse einer repräsentativen Bevölkerungsumfrage. Bertelsmann Stiftung.

Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS) (pp. 1-6). IEEE.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120.

Hardmeier, C., Costa-jussà, M. R., Webster, K., Radford, W., & Blodgett, S. L. (2021). How to Write a Bias Statement: Recommendations for Submissions to the Workshop on Gender Bias in NLP. arXiv preprint arXiv:2104.03026.

Harringer, C. (2018). „Good Bot, Bad Bot “?. Information-Wissenschaft & Praxis, 69(5-6), 257-264.

Hebert, M. H., Thorpe, C. E., & Stentz, A. (Eds.). (2012). Intelligent unmanned ground vehicles: autonomous navigation research at Carnegie Mellon (Vol. 388). Springer Science & Business Media.

Heinrichs, B. (2021). Discrimination in the age of artificial intelligence. AI & SOCIETY, 1-12.

Helbing, D., Caron, & Helbing. (2019). Towards digital enlightenment. New York, NY: Springer International Publishing.

Hersey, F. (2018). Almost 80% of Chinese concerned about AI threat to privacy, 32% already feel a threat to their work. Retrieved from https://technode.com/2018/03/02/almost-80-chinese-concerned-ai-threat-privacy-32-already-feel-threat-work/

Hielscher, S., Pies, I., Valentinov, V., & Chatalova, L. (2016). Rationalizing the GMO debate: the ordonomic approach to addressing agricultural myths. International journal of environmental research and public health, 13(5), 476.

Hobbes, T. (1655). De Corpore (p. 3.) (translated from Latin) Retrieved from https://plato.stanford.edu/entries/hobbes/

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society, 22(7), 900-915.

Hofstetter, Y. (2018). Neue Welt. Macht. Neue Menschen. Wie die Digitalisierung das Menschenbild verändert. G. Küenzlen, S. Haring-Mosbacher und P. Diehl,(Schriftenreihe/Bundeszentrale für Politische Bildung, Bd. 10247, 135–150). Bonn: bpb Bundeszentrale für Politische Bildung.

Horsley, J. P. (2021). How will China's privacy law apply to the Chinese state?. Retrieved from https://www.newamerica.org/cybersecurity-initiative/digichina/blog/how-will-chinas-privacy-law-apply-to-the-chinese-state/

Hui, L. & Tse, B. (2021). AI Governance in 2020. A year in review: Observations from 52 global experts. Retrieved from https://www.aigovernancereview.com/static/AI-Governance-in-2020-ffa2e9c4e0ec4ca3706455e0f35d5ab5.pdf

Hupperich, T., Tatang, D., Wilkop, N., & Holz, T. (2018). An empirical study on online price differentiation. In Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy (pp. 76-83).

I-AIIG (2020). The Institute for AI International Governance of Tsinghua University (I-AIIG). Retrieved from http://aiig.tsinghua.edu.cn/en/About/Overview.htm

I-AIIG (2021). The Institute for AI International Governance of Tsinghua University (I-AIIG). Retrieved from http://aiig.tsinghua.edu.cn/__local/1/50/C9/8E494613FCAD71C723A6D828519_A50F6E54_6DB4E.pdf

IBM (2017). Transparency and Trust in the Cognitive Era. Retrieved from https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/

IBM (2018a). AI Fairness 360. Retrieved from https://developer.ibm.com/open/projects/ai-fairness-360/

IBM (2018a). Introducing AI Fairness 360. Retrieved from https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/

IDC (2021). International Data Corporation. IDC Forecasts Improved Growth for Global AI Market in 2021. Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS47482321

IMF (2021). International Monetary Fund. World Economic Outlook Database. Retrieved from https://www.imf.org/en/Publications/WEO/weo-database/2021/April/weo-report?c=924,&s=NGDP_R,NGDP_RPCH,NGDP,NGDPD,PPPGDP,NGDP_D,NGDPRPC,NGDPRPPPPC,NGDPPC,NGDPDPC,PPPPC,PPPSH,&sy=2019&ey=2026&ssm=0&scsm=1&scc=0&ssd=1&ssc=0&sic=0&sort=country&ds=.&br=1

INCoDe.2030 (2021). AI Portugal 2013. Retrieved from https://www.incode2030.gov.pt/sites/default/files/julho_incode_brochura.pdf

Intel (2017). Artificial Intelligence The Public Policy Opportunity. Retrieved from https://blogs.intel.com/policy/files/2017/10/Intel-Artificial-Intelligence-Public-Policy-White-Paper-2017.pdf

Introna, L., & Wood, D. (2004). Picturing algorithmic surveillance: The politics of facial recognition systems. Surveillance & Society, 2(2/3), 177-198.

IPSOS (2020). Global Trends 2020. Understanding Complexity. Retrieved from https://www.ipsos.com/sites/default/files/ct/publication/documents/2020-02/ipsos-global-trends-2020-understanding-complexity.pdf

ISO (2021). ISO/IEC JTC 1/SC 42. Artificial intelligence. Retrieved from https://www.iso.org/committee/6794475.html

Isom, P. (2021). DoE Director on Agency's Plan to Advance Trustworthy AI. Retrieved from https://www.meritalk.com/articles/doe-director-on-agencys-plan-to-advance-trustworthy-ai/

Ito, J. (2016). Extended Intelligence. Joi Ito's PubPub. https://doi.org/10.21428/f875537b

Jia, H. (2020a). Research ethics: a safeguard for advanced technologies. National Science Review, 7(11), 1787-1792.

Jia, H. (2020b). Yi Zeng: promoting good governance of artificial intelligence. National Science Review, 7(12), 1954-1956.

Jing, M. (2019). China's tech billionaires back ethical rules to guide development of AI and other technologies. Retrieved from https://www.scmp.com/tech/enterprises/article/2188449/chinas-tech-billionaires-back-ethical-rules-guide-development-ai

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

Jordan, M. I. (2021). Stop Calling Everything AI, Machine-Learning Pioneer Says Michael I. Jordan explains why today's artificial-intelligence systems aren't actually intelligent. Retrieved from https://spectrum.ieee.org/stop-calling-everything-ai-machinelearning-pioneer-says

Joy, B. (2000). Why the future doesn't need us. Retrieved from https://www.wired.com/2000/04/joy-2/

Kant, I. (1786). Schriften zur Ethik und Religionsphilosophie (Vol. 4). Insel-Verlag.

Kasperkevic, J. (2015). Google says sorry for racist auto-tag in photo app. The Guardian. Retrieved from https://www.theguardian.com/technology/2015/jul/01/google-sorry-racistauto-tag-photo-app.

Katzenmeier, C. (2019). Big Data, E-Health, M-Health, KI und Robotik in der Medizin. Medizinrecht, 37(4), 259-271.

Kaushal A, Altman R, Langlotz C. (2020). Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. Jama, 324(12), 1212-1213.

Kendall, M. (2021). 10 Top Artificial Intelligence Startups and Companies in Malta (2021). Retrieved from https://beststartup.eu/10-top-artificial-intelligence-startups-and-companies-in-malta-2021/

Khan, S., Rahmani, H., Shah, S. A. A., & Bennamoun, M. (2018). A guide to convolutional neural networks for computer vision. Synthesis Lectures on Computer Vision, 8(1), 1-207.

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. Science and engineering ethics, 26(1), 89-120.

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. IEEE Transactions on Information Forensics and Security, 7(6), 1789-1801.

Knight, W. (2020). Baidu Breaks Off an AI Alliance Amid Strained US-China Ties. Retrieved from https://www.wired.com/story/baidu-breaks-ai-alliance-strained-us-china-ties/

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). A governance framework for algorithmic accountability and transparency.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. American psychologist, 70(6), 543.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the national academy of sciences, 110(15), 5802-5805.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 111(24), 8788-8790.

Krempl, S. (2020). EU-Kommission: 50 Millionen Startförderung fürs KI-Bündnis Claire. Retrieved from https://www.heise.de/newsticker/meldung/EU-Kommission-50-Millionen-Startfoerderung-fuers-KI-Buendnis-Claire-4684539.html

Kubat, M. (2017). An introduction to machine learning. Springer International Publishing AG.

Kurzweil, R. (2005). The singularity is near: When humans transcend biology. Penguin.

Lan, X. (2020) Professor Xue Lan accepts a special interview on the development and governance of artificial intelligence in the magazine "Internet Communication". Retrieved from https://mp.weixin.qq.com/s/O0G9zLuKAIzH6bHLAXeKgw

Lander, E. & Nelson, A. (2021). Americans Need a Bill of Rights for an AI-Powered World. Retrieved from https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/

Langkabel, T. (2021). Studie Machine Learning 2021. So kommt der KI- und ML-Erfolg. Retrieved from https://www.computerwoche.de/a/so-kommt-der-ki-und-ml-erfolg,3551614

Larochelle, H., & Bengio, Y. (2008, July). Classification using discriminative restricted Boltzmann machines. In Proceedings of the 25th international conference on Machine learning (pp. 536-543).

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094-1096.

Lee, A. et al. (2021). China's Draft Privacy Law Adds Platform Self-Governance, Solidifies CAC's Role. Retrieved from https://digichina.stanford.edu/work/chinas-draft-privacy-law-adds-platform-self-governance-solidifies-cacs-role/

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society, 5(1), 2053951718756684.

Lee, M. K., & Baykal, S. (2017). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing (pp. 1035-1048).

Leen, T., Spengler, S., Breckler, S. (2021). NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon. Retrieved from https://beta.nsf.gov/funding/opportunities/nsf-program-fairness-artificial-intelligence-collaboration-amazon

Lewontin (2015). Open AI: Effort to democratize artificial intelligence research?. Retrieved from https://www.csmonitor.com/Technology/2015/1214/Open-AI-Effort-to-democratize-artificial-intelligence-research

Li, R. (2019). Baidu CEO wants ethics research in AI strengthened. Retrieved from https://www.chinadaily.com.cn/a/201903/10/WS5c851998a3106c65c34edc5b.html

Lifang, Q. (2018). Xi Jinping: Promote the healthy development of my country's new generation of artificial intelligence. Retrieved from http://www.xinhuanet.com/politics/leaders/2018-10/31/c_1123643321.htm) with google translate

Lin, J. (2019). Face Recognition Landing Scene Observation Report (2019). Retrieved from http://epaper.oeeee.com/epaper/A/html/2019-12/06/content_52097.htm

Lin, P., Abney, K., & Bekey, G. A. (2012). Introduction to robot ethics. (pp. 3-15)

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.

Lloyd, K. (2018). Bias amplification in artificial intelligence systems. arXiv preprint arXiv:1809.07842.

Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., ... & Martinez, A. (2019). Talk to me: Exploring user interactions with the Amazon Alexa. Journal of Librarianship and Information Science, 51(4), 984-997.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2020). Gender bias in neural natural language processing. In Logic, Language, and Security (pp. 189-202). Springer, Cham.

Lu, X. (2020). Utilitarianism of Mill and Bentham: a comparative analysis. Frontiers in Educational Research, 3(4).

Lyon, D. (2005). Surveillance as social sorting: Computer codes and mobile bodies. In Surveillance as social sorting (pp. 27-44). Routledge.

Maas, M. M. (2019). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. Contemporary Security Policy, 40(3), 285-311.

Maltby, J., Day, L., & Macaskill, A. (2011). Differentielle Psychologie, Persönlichkeit und Intelligenz (Vol. 4050). Pearson Deutschland GmbH.

Marina, L. A., Trasnea, B., & Grigorescu, S. M. (2018). A multi-platform framework for artificial intelligence engines in automotive systems. In 2018 22nd International conference on system theory, control and computing (ICSTCC) (pp. 559-564). IEEE.

McAllister, A. (2016). Stranger than science fiction: The rise of AI interrogation in the dawn of autonomous robots and the need for an additional protocol to the UN convention against torture. Minn. L. Rev., 101, 2527.

McCarthy, J. (2007): What is Artificial Intelligence?. Retrieved from http://www-formal.stanford.edu/jmc/whatisai.pdf

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv preprint arXiv:1902.01007.

McNamara, A., Smith, J., & Murphy-Hill, E. (2018, October). Does ACM's code of ethics change ethical decision making in software development?. In Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering (pp. 729-733).

Megvii (2019). Megvii Technology Limited. p. 154. Retrieved from https://ipvm-uploads.s3.amazonaws.com/uploads/7490/c099/megvii-ipo.pdf

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Merriam-Webster (2021). Definition of ethic. Retrieved from https://www.merriam-webster.com/dictionary/ethics

Mill, J. S. (1863). Utilitarianism London: Parker. Son and Bourn.

Minsky, M., Feigenbaum, E. A., & Feldman, J. (1963). Computers and thought.

Moor, J. H. (2020). The Mature, Importance, and Difficulty of Machine Ethics. In Machine Ethics and Robot Ethics (pp. 233-236). Routledge.

Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K., & Grossman, D. (2018). Military Applications of Artificial Intelligence. Ethical Concerns in an Uncertain World. RAND Corporation, Santa Monica CA.

Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K., & Grossman, D. (2020). Military applications of artificial intelligence: ethical concerns in an uncertain world. RAND PROJECT AIR FORCE SANTA MONICA CA SANTA MONICA United States.

MOST (2019) China: AI Governance Principles Released. Retrieved from the Library of Congress. Retrieved from https://www.loc.gov/item/global-legal-monitor/2019-09-09/china-ai-governance-principles-released/.

Mozur, P. (2017). A.I. Expert at Baidu, Andrew Ng, Resigns From Chinese Search Giant. Retrieved from https://www.nytimes.com/2017/03/22/business/baidu-artificial-intelligence-andrew-ng.html

Mullen, B., & Hu, L. T. (1989). Perceptions of ingroup and outgroup variability: A meta-analytic integration. Basic and Applied Social Psychology, 10(3), 233-252.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. Political Behavior, 39(3), 629-649.

Munshi, A., & Sharma, V. (2018). Safety and ethics in biotechnology and bioengineering: What to follow and what not to. In Omics Technologies and Bio-Engineering (pp. 577-590). Academic Press.

Munshi, A., & Sharma, V. (2018). Safety and ethics in biotechnology and bioengineering: What to follow and what not to. In Omics Technologies and Bio-Engineering (pp. 577-590). Academic Press.

Murphy, S. (2021). https://www.federalregister.gov/documents/2021/10/08/2021-21975/notice-of-request-for-information-rfi-on-public-and-private-sector-uses-of-biometric-technologies

Nadella, S. (2016). Microsoft CEO Satya Nadella lays out 10 Laws of AI (and Human Behavior). Retrieved from https://www.geekwire.com/2016/microsoft-ceo-satya-nadella-10-laws-ai/

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., ... & Martino, G. D. S. (2021). Automated fact-checking for assisting human fact-checkers. arXiv preprint arXiv:2103.07769.

Nana, F. (2019). What should artificial intelligence companies do if they want to form an ethics committee?. Retrieved from http://www.bjnews.com.cn/feature/2019/07/26/608130.html

Neff, G., Nagy, P. (2016). Automation, algorithms, and politics| talking to Bots: Symbiotic agency and the case of Tay. International Journal of Communication 10: 4915–4931

Neuhäuser, C. (2012). Künstliche Intelligenz und ihr moralischer Standpunkt (pp. 23-42). Nomos Verlagsgesellschaft mbH & Co. KG.

Ng (2017). Opening a new chapter of my work in AI. Retrieved from https://medium.com/@andrewng/opening-a-new-chapter-of-my-work-in-ai-c6a4d1595d7b#.shvlsibzs

Nilsson, N. J. (2010). The quest for artificial intelligence. A history of ideas and achievements. Retrieved from https://ai.stanford.edu/~nilsson/QAI/qai.pdf.

NIST (2021a). Summary Analysis of Responses to the NIST Artificial Intelligence Risk Management Framework (AI RMF) - Request for Information (RFI) National Institute of Standards and Technology (NIST)

NIST (2021b). Response of Microsoft Corporation to NIST RFI on an Artificial Intelligence Risk Management Framework. Retrieved from https://www.nist.gov/system/files/documents/2021/09/16/ai-rmf-rfi-0088.pdf

NIST (2021c). National Institute of Standards and Technology. Comments Received for RFI on Artificial Intelligence Risk Management Framework. Retrieved from https://www.nist.gov/itl/ai-risk-management-framework/comments-received-rfi-artificial-intelligence-risk-management

NLPR (2018). National Laboratory of Pattern Recognition. "Artificial Intelligence Ethics Research Project Launch Conference and Artificial Intelligence Ethics, Policy and Law Forum" held in Beijing. Retrieved from http://www.nlpr.ia.ac.cn/cn/news/664.html

NSCAI (2019). National Security Commission on Artificial Intelligence. Retrieved from https://epic.org/wp-content/uploads/foia/epic-v-ai-commission/AI-Commission-Interim-Report-Nov-2019.pdf

OECD (2019). Organisation for Economic Co-operation and Development. Forty-two countries adopt new OECD Principles on Artificial Intelligence. Retrieved from https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm.

OECD (2020). Organisation for Economic Co-operation and Development. OECD to host Secretariat of new Global Partnership on Artificial Intelligence. Retrieved from https://www.oecd.org/going-digital/ai/oecd-to-host-secretariat-of-new-global-partnership-on-artificial-intelligence.htm.

ÓhÉigeartaigh, S. S., Whittlestone, J., Liu, Y., Zeng, Y., & Liu, Z. (2020). Overcoming barriers to cross-cultural cooperation in AI ethics and governance. Philosophy & technology, 33(4), 571-593.

Olson, P. (2018). The algorithm that helped google translate become sexist. Retrieved from https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/?sh=7fc28bd7daa2

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

OP (2018) Commitments and principles. Retrieved from https://www.op.fi/op-financial-group/corporate-social-responsibility/commitments-and-principles

OpenAI (2018). OpenAI Charter. Retrieved from https://openai.com/charter/

PAI (2016). About the Partnership on AI. Retrieved from https://partnershiponai.org/about/

PAI (2018). Introducing Our First Chinese Member to the Partnership on AI. Retrieved from https://www.partnershiponai.org/introducing-our-first-chinese-member-to-the-partnership-on-ai/

PAI (2021). About the Partners of the Partnership on AI. Retrieved from https://partnershiponai.org/partners/

Passe, J., Drake, C., & Mayger, L. (2018). Homophily, echo chambers, & selective exposure in social networks: What should civic educators do?. The Journal of Social Studies Research, 42(3), 261-271.

Paul, R., & Elder, L. (2019). The miniature guide to critical thinking concepts and tools. Rowman & Littlefield.

Piaget, J. (2000). Psychologie der intelligenz. Klett-Cotta.

Pichai, S. (2018). Google CEO: A.I. is more important than fire or electricity. Retrieved from https://www.cnbc.com/2018/02/01/google-ceo-sundar-pichai-ai-is-more-important-than-fire-electricity.html

Pinker, S. (2005). So how does the mind work?. Mind & Language, 20(1), 1-24.

Poel, M., Meyer, E. T., & Schroeder, R. (2018). Big data for policymaking: Great expectations, but with limited progress?. Policy & Internet, 10(3), 347-367.

Porter, J. (2019). Another convincing deepfake app goes viral prompting immediate privacy backlash. Retrieved from https://www.theverge.com/2019/9/2/20844338/zao-deepfake-app-movie-tv-show-face-replace-privacy-policy-concerns

Porter, T. M., (1996). Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. ISIS-International Review Devoted to the History of Science and its Cultural Influence, 87(3), 519-520.

Powers, T. M. (2006). Prospects for a Kantian machine. IEEE Intelligent Systems, 21(4), 46-51.

Prates, M., Avelar, P., & Lamb, L. C. (2018). On quantifying and understanding the role of ethics in AI research: A historical account of flagship conferences and journals. arXiv preprint arXiv:1809.08328.

Qian, Y., Muaz, U., Zhang, B., & Hyun, J. W. (2019). Reducing gender bias in word-level language models with a gender-equalizing loss function. arXiv preprint arXiv:1905.12801.

Rahmani, S., Mousavi, S. M., & Kamali, M. J. (2011). Modeling of road-traffic noise with the use of genetic algorithm. Applied Soft Computing, 11(1), 1008-1013.

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. Ethics and Information Technology, 20(1), 5-14.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. Nature, 568(7753), 477-486.

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 429-435).

Rao, A., Verweij, G., Cameron, E. (2017). Sizing the prize What's the real value of AI for your business and how can you capitalise?. Retrieved from https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf

Resnik, D. B. (2011). What is ethics in research and why is it important. National Institute of Environmental Health Sciences, 1(10), 49–70.

Rich, A. S., & Gureckis, T. M. (2019). Lessons for artificial intelligence from the study of natural stupidity. Nature Machine Intelligence, 1(4), 174-180.

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. Big Data & Society, 3(1), 2053951716649398.

Rome Call (2020). renAIssance. The Pontifical Academy for Life organised the congress “RenAIssance. For a Human-centric Artificial Intelligence” in Rome on the 28th February 2020, that culminated in the signature of the Rome Call for AI Ethics. Retrieved from https://romecall.org/

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Rossi, F. (2018) Everyday Ethics for Artificial Intelligence: a guide for Designers and Developers. https://futurium.ec.europa.eu/sk/european-ai-alliance/blog/everyday-ethics-artificial-intelligence-guide-designers-and-developers?language=pt-pt

Russell, S. (2021). ‘Yeah, we’re spooked’: AI starting to have big real-world impact, says expert. Retrieved from https://www.theguardian.com/technology/2021/oct/29/yeah-were-spooked-ai-starting-to-have-big-real-world-impact-says-expert

Russell, S., & Norvig, P. (1995). A modern, agent-oriented approach to introductory artificial intelligence. ACM SIGART Bulletin, 6(2), 24-26.

Russell, S., & Norvig, P. (2003). Artificial Intelligence: A Modern Approach.

Sabbagh, D. (2021). Facebook trained its AI to block violent live streams after Christchurch attacks. Retrieved from https://www.theguardian.com/technology/2021/oct/29/facebook-trained-its-ai-to-block-violent-live-streams-after-christchurch-attacks

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Sandler, R., Basl, J., Tiell. S. C. (2019). Data and AI Ethics Committees. Technical Report. Accenture. Retrieved from https://www.accenture.com/us-en/insights/software-platforms/building-data-ai-ethics-committees

SAP (2018a). SAP Becomes First European Tech Company to Create Ethics Advisory Panel for Artificial Intelligence. Retrieved from https://news.sap.com/2018/09/sap-first-european-tech-company-ai-ethics-advisory-panel/

SAP (2018a). SAP’s Guiding Principles for Artificial Intelligence. Retrieved from https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/

SAP (2021). Solve complex societal problems with artificial intelligence. Retrieved from https://www.sap.com/products/artificial-intelligence/ai-ethics.html

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. arXiv preprint arXiv:2104.06001.

Schiebinger, L. (2014). Scientific research must take gender into account. Nature News, 507(7490), 9.

Schmidt, E. (2021a). Misinformation Is About to Get So Much Worse. https://www.theatlantic.com/technology/archive/2021/09/eric-schmidt-artificial-intelligence-misinformation/620218/

Schmidt, E. (2021b). U.S. is ‘not prepared to defend or compete in the A.I. era,’ says expert group chaired by Eric Schmidt. Retrieved from https://www.cnbc.com/2021/03/02/us-not-prepared-to-defend-or-compete-in-ai-era-says-eric-schmidt-group.html

Schmidt, E., Work, B., Catz, S., Chien, S., Darby, C., Ford, K., ... & Moore, A. (2021). National Security Commission on Artificial Intelligence (AI). National Security Commission on Artificial Intellegence.

Schuster, T., Shah, D. J., Yeo, Y. J. S., Filizzola, D., Santus, E., & Barzilay, R. (2019). Towards debiasing fact verification models. arXiv preprint arXiv:1908.05267.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and brain sciences, 3(3), 417-424.

Shah, D., Schwartz, H. A., & Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview. arXiv preprint arXiv:1912.11078.

Shaohua, H. (2019). The Chinese Academy of Social Sciences holds a seminar on "The Society, Ethics and Future of Artificial Intelligence". Retrieved from https://www.thepaper.cn/newsDetail_forward_3381346

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. Ethics and information technology, 14(1), 27-40.

Sharkey, N. (2016). Staying in the loop: human supervisory control of weapons. Autonomous weapons systems: Law, ethics, policy, 23-38.

Shearer, E., Pasquarelli, W. & Stirling, R. (2020). Government AI Readiness Index 2020. Retrieved from https://static1.squarespace.com/static/58b2e92c1e5b6c828058484e/t/5f7747f29ca3c20ecb598f7c/1601653137399/AI+Readiness+Report.pdf

Shen, Y. & Wang, Q. (2020). Ethics Arguments on AI in Education: An Analysis of the EU's Ethics Guidelines for Trustworthy AI from an Educational Perspective)." In: Peking University Education Review (2019) 17 (4): 18-34

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: ethics of AI and ethical AI. Journal of Database Management (JDM), 31(2), 74-87.

Simmonds, L. (2021). Croatian Artificial Intelligence Ecosystem to be Mapped Once Again. Retrieved from https://www.total-croatia-news.com/made-in-croatia/56262-croatian-artificial-intelligence-ecosystem.

Simon, H. A. (1944). Decision-making and administrative organization. Public Administration Review, 4(1), 16-30.

Simons, H. (2009). Case study research in practice. SAGE publications.

Smith, A. & Anderson, M. (2017). Automation in Everyday Life. Retrieved from https://www.pewresearch.org/internet/2017/10/04/automation-in-everyday-life/

Smith, D. (2013). IBM's Watson Gets A 'Swear Filter' After Learning The Urban Dictionary. Retrieved from https://www.ibtimes.com/ibms-watson-gets-swear-filter-after-learning-urban-dictionary-1007734

SPC (2015). Supreme People's Court. Introduction. Retrieved from http://english.court.gov.cn/2015-07/16/content_21299713.htm

Stake, R. E. (1995). The art of case study research. sage.

State Council (2017). State Council on Printing and Distributing Notice on the development plan of the new generation of artificial intelligence. Retrieved from https://chinacopyrightandmedia.wordpress.com/2017/07/20/a-next-generation-artificial-intelligence-development-plan/

STOA (2021). Scientific Foresight Unit (STOA) PE 690.039 Tackling deepfakes in European policy. Retrieved from https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf

Stone, L. (2020). Baidu leaves 'Partnership on AI' as US-China relations sour. Retrieved from https://aibusiness.com/document.asp?doc_id=761881

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

Sullivan, R. J. (1994). An introduction to Kant's ethics. Cambridge University Press.

Taebi, B., Roeser, S., & van de Poel, I. (2012). The ethics of nuclear power: Social experiments, intergenerational justice, and emotions. Energy Policy, 51, 202-206.

TAIL (2019). Tencent AI Lab. Technological ethics in the intelligent age-reshaping the trust of the digital society. Retrieved from https://tisi.org/10890

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. Big Data & Society, 4(2), 2053951717736335.

Technavio (2021). Artificial Intelligence (AI) Chips Market to grow by USD 73.49 billion|Technavio. Retrieved from https://www.prnewswire.com/news-releases/artificial-intelligence-ai-chips-market-to-grow-by-usd-73-49-billiontechnavio-301346619.html

Telefónica (2018) AI Principles of Telefonica. Retrieved from https://www.telefonica.com/wp-content/uploads/sites/7/2021/11/principios-ai-eng-2018.pdf

Tencent (2019). China's Tencent advocates "AI for Good" at AI Everything Summit in Dubai. Retrieved from http://www.chinadaily.com.cn/a/201905/02/WS5ccaca1ba3104842260b98a0.html

The Economist (2017). China may match or beat America in AI. Retrieved from https://www.economist.com/news/business/21725018-its-deep-pool-data-may-let-it-lead-artificial-intelligence-china-may-match-or-beat-america

Thiel, P. (2019). Report: Palantir took over Project Maven, the military AI program too unethical for Google. Retrieved from https://thenextweb.com/news/report-palantir-took-over-project-maven-the-military-ai-program-too-unethical-for-google

Thierer, A., O'Sullivan Castillo, A., & Russell, R. (2017). Artificial intelligence and public policy. Mercatus research. Mercatus Center at George Mason University.

TietoEVRY (2018). Tieto strengthens commitment to ethical use of AI. Retrieved from https://www.tietoevry.com/en/newsroom/all-news-and-releases/press-releases/2018/10/tieto-strengthens-commitment-to-ethical-use-of-ai/

Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. PLoS one, 13(9), e0203958.

Tracxn (2021a). AI Start Ups United States. Retrieved from https://tracxn.com/explore/Artificial-Intelligence-Startups-in-United-States

Tracxn (2021b). AI Start Ups China. Retrieved from https://tracxn.com/explore/Artificial-Intelligence-Startups-in-China

Tracxn (2021c). AI Start Ups EU. Retrieved from https://tracxn.com/explore/

Trappl, R. (1986). Impacts of artificial intelligence. North Holland Publishing Co.

TRI (2017) Tencent Research Institute et al., 2017; translated by Ding 2018

Turek, M. (2021). Explainable Artificial Intelligence (XAI). Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence

Turing, A. M. (1950): Computing machinery and intelligence. In: Mind, New Series, 1950, 59. Jg., Nr. 236, S. 433-460.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive psychology, 5(2), 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. science, 185(4157), 1124-1131.

UN (2020). High-level Panel on Digital Cooperation. Retrieved from https://www.un.org/en/sg-digital-cooperation-panel

UNESCO (2019). International Conference on Artificial Intelligence and Education. Retrieved from https://en.unesco.org/themes/ict-education/ai-education-conference-2019

Unity (2018). Introducing Unity's Guiding Principles for Ethical AI. Retrieved from https://blog.unity.com/technology/introducing-unitys-guiding-principles-for-ethical-ai

Vaes, J., Bain, P. G., & Bastian, B. (2014). Embracing humanity in the face of death: why do existential concerns moderate ingroup humanization?. The Journal of social psychology, 154(6), 537-545.

Vakkuri, V., & Abrahamsson, P. (2018, June). The key concepts of ethics of artificial intelligence. In 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC) (pp. 1-6). IEEE.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2), 2053951717743530.

Verba, S. (2006). Fairness, equality, and democracy: three big words. Social Research, 499-540.

Vought, R. T. (2020). Guidance for Regulation of Artificial Intelligence Applications. Retrieved from https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf

Wang, J (2021). Notice on Soliciting Opinions on the Draft of the National Standard "Information Security Technology Face Recognition Data Security Requirements". Secretariat of the National Information Security Standardization Technical Committee. Retrieved from https://www.tc260.org.cn/front/bzzqyjDetail.html?id=20210423182442&normid=20201104200034&recode_id=41855

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology, 114(2), 246.

Ward, A. F., Duke, K., Gneezy, A., & Bos, M. W. (2017). Brain drain: The mere presence of one's own smartphone reduces available cognitive capacity. Journal of the Association for Consumer Research, 2(2), 140-154.

Webb, P. T. (2014). How to do your case study: a guide for students and researchers.

Webster, G. (2019). Translation: Chinese AI Alliance Drafts Self-Discipline 'Joint Pledge'. Retrieved from https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-ai-alliance-drafts-self-discipline-joint-pledge/

Wei, H. (2020). Shanghai introduces AI investment consortium. Retrieved from http://english.www.gov.cn/news/topnews/202007/11/content_WS5f0916f0c6d06c4091250be7.html

Wei, l. D. (2021). China Online Medicine Shares Tumble as Beijing Clarifies Rules. Retrieved from https://finance.yahoo.com/news/china-online-medicine-shares-tumble-100956116.html

Weiss, G. (1999). Multiagent systems: a modern approach to distributed artificial intelligence. MIT press.

Weiwen, D. (2019). Humans and machines must understand each other's defects and co-evolve. Retrieved from https://homest.org.cn/article/detail?id=509997

Weyerer, C. J., & Langer, F. P. (2019). Garbage in, garbage out: The vicious cycle of AI-based discrimination in the public sector. In Proceedings of the 20th Annual International Conference on Digital Government Research (pp. 509-511).

Weyerer, J. C., & Langer, P. F. (2020). Bias and Discrimination in Artificial Intelligence: Emergence and Impact in E-Business. In Interdisciplinary Approaches to Digital Transformation and Innovation (pp. 256-283). IGI Global.

White, M. C. (2012). Orbits Shows Higher Prices to Mac Users. Time. Retrieved from https://www.business.time.com/2012/06/26/orbitz-shows-higherprices-to-mac-users.

Wiggers, K. (2019). CB insights: Here are the top 100 AI companies in the world. Retrieved from https://venturebeat.com/2019/02/06/cb-insights-here-are-the-top-100-ai-companies-in-the-world/.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. International Journal of Public Administration, 42(7), 596-615.

Woolley, S., & Joseff, K. (2018). Computational Propaganda, Jewish-Americans and the 2018 Midterms: The Amplification of Anti-Semitic Harassment Online. October. The Anti-Defamation League.

Worldometer (2021). Worldometer Chinese Population. Retrieved from https://www.worldometers.info/world-population/china-population/

WSP (2021). Washington Post. Retrieved from https://www.washingtonpost.com/wp-srv/world/countries/china.html

Xiang, X. (2020). Translated from Chinese: The Formation and Teaching of the EU'Strategy for Medical AI.

Yeh, T. M. (2017). Designing a moral compass for the future of computer vision using speculative analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 64-73).

Yin, R. K. (2003). Design and methods. Case study research, 3(9.2).

Yin, R. K. (2009). Case study research: Design and methods (Vol. 5). sage.

Yin, R. K. (2011). Applications of case study research. sage.

Ying, F. (2019). Understanding the AI Challenge to Humanity. Retrieved from https://www.chinausfocus.com/foreign-policy/understanding-the-ai-challenge-to-humanity

Yong, H. (2017) 3TH1CS: die Ethik der digitalen Zeit. iRights Media. (pp. 189-209)

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. arXiv preprint arXiv:1812.02953.

Yuan, L. (2018). How Cheap Labor Drives China's A.I. Ambitions. The New York Times. Retrieved from https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html

Zanzotto, F. M. (2019). Human-in-the-loop Artificial Intelligence. Journal of Artificial Intelligence Research, 64, 243-252.

Zeng, Y. (2018b). Harmonious Artificial Intelligence Principles. Introduction to the Harmonious Artificial Intelligence Principles: Vision and Mission. Retrieved from http://harmonious-ai.org/

Zeng, Y. (2019). The Official Launch of ChinUK Centre for AI Ethics and Governance. Retrieved from https://ai-ethics-and-governance.institute/2019/11/12/the-official-launch-of-chinuk-centre-for-ai-ethics-and-governance/)

Zeng, Y. (2021). Global Cooperation on Artificial Intelligence is not a Zero-Sum Game. Retrieved from https://ai-ethics-and-governance.institute/2021/02/03/global-cooperation-on-artificial-intelligence-is-not-a-zero-sum-game/

Zeng, Y., Lu, E., & Huangfu, C. (2018a). Linking artificial intelligence principles. arXiv preprint arXiv:1812.04814.

Zeng, Y., Lu, E., Sun, Y., & Tian, R. (2019). Responsible facial recognition and beyond. arXiv preprint arXiv:1909.12935.

Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends. Available at SSRN 3312874.

Zhang, B., & Dafoe, A. (2020, February). US public opinion on the governance of artificial intelligence. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 187-193).

Zhang, D. Mishra, S., Brynjolfsson, E., Etchemendy, J. (2021). The AI Index 2021 Annual Report.

Zhong, D. (2019). Notice on Issuing the "Regulations on the Administration of Network Audio and Video Information Services". Retrieved from http://www.nrta.gov.cn/art/2019/11/29/art_113_48908.html