KADİR HAS ÜNİVERSİTESİ
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF COMMUNICATION SCIENCES

# HOW DOES ARTIFICIAL INTELLIGENCE EVALUATE? AN EXAMINATION INTO THE DEPLOYMENT OF ARTIFICIAL INTELLIGENCE IN THE TURKISH ECOSYSTEM

ŞEYDA TUĞGEN GÜMÜŞAY

SUPERVISOR: ASSOC. PROF. LEVENT SOYSAL

MASTER'S THESIS

ISTANBUL, JULY 2021

# HOW DOES ARTIFICIAL INTELLIGENCE EVALUATE? AN EXAMINATION INTO THE DEPLOYMENT OF ARTIFICIAL INTELLIGENCE IN THE TURKISH ECOSYSTEM

ŞEYDA TUĞGEN GÜMÜŞAY

SUPERVISOR: ASSOC. PROF. LEVENT SOYSAL

MASTER'S THESIS

Submitted to the School of Graduate Studies of Kadir Has University in partial fulfillment of the requirements for the degree of Master's in the Programme of New Media under the Department of Communication Studies.

ISTANBUL, OCTOBER 2021

## ACCEPTANCE AND APPROVAL

This work entitled **How Does Artificial Intelligence Assess Humans: The Relationship Between Artificial Intelligence and Societal Judgements?** prepared by **SEYDA TUGGEN GUMUSAY** has been judged to be successful at the defense exam held on **28.10.2021** and accepted by our jury as **MASTERS'S THESIS.**

| | |
|---|---|
| Associate Proffesor Levent Soysal) (Supervisor) | Kadir Has University |
| Associate Proffesor İrem İnceoğlu | Kadir Has University |
| Associate Proffesor Özgür Narin | Ordu University |

I certify that the above signatures belong to the faculty members named above.

İMZA
Müdür
Lisansüstü Eğitim Enstitüsü
ONAY TARİHİ: Gün/Ay/Yıl

I, SEYDA TUGGEN GUMUSAY OF THE CANDIDATE,

Hereby declare that this master's thesis is my own original work and that due references have been appropriately provided on all supporting literature and resources.

ŞEYDA TUĞGEN GÜMÜŞAY

_____

TARİH VE İMZA

# TEŞEKKÜR

I would like, firstly, to thank my dear friends, Büşra Sağlam, Yiğit Bahadır Kaya, Can Alkan, Talha Yılmaz, Ersin Güney, Hümeyra Nur Hatipoğlu and Ceren Güler, for their support of my writing process, being always there for me when my motivation decreased and giving me courage.

I also would like to thank TÜBİTAK BİDEB for their financial support of my education.

I would like to thank Esra Soybir on behalf of many people who were with me in every step I took, knowing that if I tried to list their names, I would leave one missing.

I would also like to express my greatest thanks to my advisor, Assoc. Dr. Levent Soysal, for his support and trust in my academic education.

HOW DOES ARTIFICIAL INTELLIGENCE EVALUATE? AN EXAMINATION INTO THE DEPLOYMENT OF ARTIFICIAL INTELLIGENCE IN THE TURKISH ECOSYSTEM

# ABSTRACT

With the development of artificial intelligence, machine-based applications are entering people's lives more and more every day. With artificial intelligence, especially as a decision-making mechanism, research conducted in recent years shows that artificial intelligence is biased. This study examines the bias and discriminatory behavior of artificial intelligence in detail. Inherited bias, data-based factors, and technical reasons that cause artificial intelligence bias are explained, and a general framework for the reasons for bias is shown. Models are presented for the discussion ground and the social shaping of artificial intelligence and technology in the context of "biasedness, discrimination, companies, and the system" is examined. The underlying reasons for the bias of artificial intelligence have been investigated. For this, Turkey's ecosystem is taken as a case study, and companies' attitudes towards the bias of AI are examined through interviews. Within the framework of social shaping theory, the structural connection between the bias of artificial intelligence and the economic structure and capitalist system has been pointed out.

**Keywords:** Artificial Intelligence, bias, discrimination, capitalism, social shaping of technology, technological determinism

YAPAY ZEKA NASIL DEĞERLENDİRİR? TÜRKİYE EKOSİSTEMİNDE YAPAY ZEKANIN YERLEŞİMİ ÜZERİNE BİR İNCELEME

# ÖZET

Yapay zekanın gelişmesi ile birlikte makine temelli uygulamalar her geçen gün insanların hayatına daha fazla girmektedir. Yapay zekanın özellikle karar verme mekanizması olarak kullanılması ile son yıllarda yapılan araştırmalar yapay zekanın taraflı olduğunu göstermektedir. Bu çalışmada yapay zekanın taraflılığı ve ayrımcı davranışları detaylıca incelemektedir. Yapay zekanın taraflılığına sebep olan toplumsal aktarımlar, veri bazlı etkenler ve teknik sebepler anlatılmış, taraflılığın sebeplerine ilişkin genel bir çerçeve gösterilmiştir. Tartışma zemini için modeler sunulmuş ve "taraflılık, ayrımcılık, şirketler ve sistem" bağlamında yapay zeka ve teknolojinin toplumsal olarak şekillendirilişi incelenmiştir. Türkiye ekosistemi üzerinden şirketlerin yapay zekaya ilişkin tutumları görüşmeler aracılığıyla belirlenerek yapay zekanın taraflılığının özünde yatan sebepler araştırılmıştır. Teknolojinin sosyal inşası çerçevesinde yapay zeka taraflılığının ekonomik yapıyla ve kapitalist sistemle yapısal bağı işaret edilmiştir.

**Anahtar Sözcükler:** Yapay zeka, taraflılık, ayrımcılık, kapitalizm, teknolojinin sosyal inşası, teknolojik determinizm

TABLE OF CONTENTS

# ABBREVIATION LIST

| | |
|---|---|
| AI | Artificial Intelligence |
| FBTD | From Bias to Discrimination |
| HLEG ACAI | High-Level Expert on Artificial Intelligence |
| ML | Machine Learning |
| NLP | Natural Language Processing |
| RBTD | Re-categorized Bias to Discrimination |
| RL | Reinforcement Learning |
| RPA | Robot Process Automation |
| SST | Social Shaping of Technology |
| TRAI | Turkey Artificial Intelligence Initiative |

# LIST OF TABLES

# LIST OF FIGURES

# 1. INTRODUCTION

*"The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else (Yudkowsky 2008, 333)"*

Throughout history, technological developments have always been part of humans' lives. In the last fifty years, artificial intelligence (AI) has become one of the most discussed topics in technology. Although AI is not new today, and studies and research focus on practical and usable sides of AI, in recent years, discussions on social-related aspects have progressed significantly. Algorithmic decisions frequently perform in many fields, such as healthcare, education, banking, e-commerce. For example, deep learning algorithms can treat anonymized electronic health reports and flag potential dangers, to which clinicians are then immediately able to react. With banks moving towards mobile payments to offer a seamless and fast customer experience, payment services based on machine learning algorithms verify and identify credit fraud in real-time. Likewise, automated data credibility evaluation methods are used for speed approval, verification, or detection of unusual activities by insurance accompanies.

The technical developments in AI and implementing AI in daily lives like security, recognition, or tracking services lead to theoretical discussions for AI. With the latest studies, a recent problem has emerged: discrimination of AI. Researchs presents that many algorithmic decisions exhibit biased results. In 2016, Tay, a Microsoft Twitter chatbot, has become a popular topic because of her behaviors, which are sexist and racist. The chatbot was designed to "engage and entertain people where they talk to each other online through casual and playful conversation" (Tennery and Cherelus 2016, Reuters, date 24.03.2016), and targeted American 18 to 24 years old primary social media users, however more she talks with the people, her conversation has extended to racist, inflammatory and political statements (Hunt 2016, The Guardian, date 24.03.2016). The statement of Tay as "Hitler was right" and "I (expletive) hate feminists, and they should all die and burn in hell." After her tweets, Microsoft has ended the project and deleted all the tweets; however, it has become publicly appear, AI can exhibit discriminatory

behaviors[i]. After Tay's decisions, the search for artificial intelligence has increased. And in the last years, "Allegations of racism and sexism, has permeated the conversation as stories surface about search engines delivering job postings for well-paying technical jobs to men and not women, or providing arrest mugshots when keywords such as 'black teenagers' are entered is seen" (Howard and Borenstein 2018). Before these studies, AI was equal to some algorithms and statistical information. For many, AI was giving calculated and unbiased results. However, studies put different outcomes.

In 2021, I read the article titled "AI finally closing in on human intelligence?" written by John Thornhill, and he stated that:

> "AI might only multiply many of the problems we confront today: the excessive concentration of corporate power as private companies increasingly assume the functions once exercised by nation-states; the further widening of economic inequality and the narrowing of opportunity; the spread of misinformation and the erosion of democracy."(Thornhill 2021, Wired, date 11.11.2020)

Within thinking in terms of AI's potential harms and usage of AI in many fields, the bias and discrimination of AI, and the reason behind it, becomes more critical for me. Debates on the "representative harm and allocative harm" of AI's bias also reinforce the idea of examining the bias of AI (Crawford 2021).

With all this information, many questions come to mind. Did human intelligence create a technology that is more improved, less biased, more rational than human intelligence itself? Or the constraint on AI is a barrier to AI's perception of particular identities of humans? Can AI strengthen the existed bias and marginalize particularities? Is AI more rational than human, or rationality of AI depends on utility? These are important questions to ask.

By seeing these questions, the bias of discrimination of AI becomes crucial. For that, the study aims to describe the relationship between bias, discrimination, and society in the context of economic structure. This study aims to show relations between bias and discrimination for artificial intelligence and seeks to answer questions about how societal bias related to artificial intelligence and how society's structure affects the occurrence of discrimination. It also figures out attitudes of the companies. Within this context, these questions also need answers: do artificial intelligence-based companies realize the bias,

or to what extent they can detect bias, what is the priority for companies in the context of bias or is bias priority, and how they try to prevent discriminatory behaviors of artificial intelligence.

In this thesis, I argue data contains the bias of society and transmits bias to AI, and I hypothesize that bias is a structural condition linked to the socio-economic system, not "directly" to data or AI itself. So, the thesis argues that the bias of artificial intelligence is inherited bias that comes from society itself. The community's economic structure reinforces companies to overlook prejudice or discrimination. The market competition policies and profits created by capitalism and the superstructure of capitalism, including culture, technology, daily life, etc., are strongly in relation to bias. So, we must discuss artificial intelligence bias in terms of socio-economic systems and capitalism. So, the thesis contributes the literature from the social sciences' perspective and aims to provide a new perspective for debate on bias.

The study based on the interpretation and context, so the qualitative method is the proper one. Showing the inductive relationship between theory and method is appropriate for the study because the central approach is to examine the "bias, AI, company and system" relationship. Therefore, the analysis depends on epistemologically interpretive and ontologically constructionist approaches.

For study, I chose Turkey's ecosystem as a case study. I carried qualitative research in the research to test the hypothesis because the hypothesis proposes a structure that constructs a distinct and comprehensive picture, as well as framework, depends on context, interpretation, and reflections. There is an inductive relationship between theory and method. Because of my central intention is to explain the relations between artificial intelligence and institutional policies, epistemologically it is interpretive and ontologically constructionist.

To conduct research, I collected primary data through interviews. I chose interview because there are informants who have information about the subject that I inquire. "Design in qualitative interviewing is iterative. That means that each time you repeat the

basic process of gathering information, analyzing it, winnowing it, and testing it, you come closer to a clear and convincing model of the phenomenon you are studying" (Herbert and Rubin 1995, 43). My intention is to understand constraints, cognitive beliefs, perceptions of bias, discrimination and AI; therefore, the interview is the proper way. The plus of interview method is "flexible, adaptive, and responsive to the experiences and utterances of the informant" (Baxter and Babbie 2003, 344) which reveals the themes of study.

For the study, I followed seven stages of Kvale as "thematizing, designing, interviewing, transcribing, analyzing, verifying, and reporting" is followed (1996, 88) As a method of interview, semi-structured interview is chosen. Semi structured interview "characterized by substantial freedom on the part of informant" (Baxter and Babbie 2003, 329), which is important to understand different motives and approaches of interviewees. Semi structured interview is also helpful because a wide range of artificial intelligence literature and the current discussion on too many topics related to AI, few basic open-ended questions that do not control the interview but guide the interview in the axis of "context" are determined. A semi-structured interview reveals repetitive and common expressions, frequencies of answers, and provides more opportunities for contextual analysis. I chose descriptive and structural questions for the interview questions. Questions can be seen in Appendix A.

In this thesis, in next chapter, to comprehend artificial intelligence, the historical development of AI has given. Then, chapter provides definitions and different approaches to AI and presents several schemas for AI, subsets of AI, and Machine Learning (ML).

In the third chapter, a framework to understand what means bias and discrimination for AI is defined. The literature on the bias of AI is critical to see various approaches. Different studies on discrimination of AI are given and different bias types are defined.

In the fourth chapter, I proposed two models to clarify relationship between society, bias, and discrimination. The logic of the models depends on social shaping of technology theory. So, in third chapter, theoretical approaches to technology are stated.

Following a brief theoretical framework, in the fifth chapter, the research method is presented. After providing a general outlook for Turkey's AI ecosystem, I introduced the interviewers. For research, seven people who work in the AI sector are chosen. As a research method, a semi-structured interview has been selected. Questions of interviews can be found in Appendix A. In the third chapter, job definition of interviewers and how they use artificial intelligence is presented. After it, the issues that interviewers point to have been explained in detail. Several topics, such as dataset bias, inherited bias, the importance of performance in AI, have been stated. The questions of who is responsible for the bias of AI and how companies prevent discrimination are answered.

In the last chapter, a brief conclusion of the study is given and a general discussion on bias, society, and system has been introduced for further studies. The aim of the chapter is to show the present approach of the bias of AI and contribute to the field.

# 2. FRAMING ARTIFICIAL INTELLIGENCE

Artificial Intelligence is one of the most innovative progress in the 21st century. It has taken significant consideration from the news, magazines, scholars, or media sector. As this technology advances, the study on AI has extended. Various subjects such as mind theories and AI, media affects, gender studies, ethical dilemmas have become a widespread interest for scholars. Besides the academic concerns, the production process, the means for consumption advertising, and people's daily lives have also affected. Stanford's Professor of Computer Science Fei-Fei Li asserted AI would become an essential part of the fourth industrial revolution, and it would influence all parts of people's lives (Hempel 2017, Wired, date 04.05.2017). As a technological development, attention for AI is still going on in both applications to the daily lives of human and academic debates. In this chapter, I give the historical background of AI. After it, AI and machine learning are described, which are important to grasp the theoretical references for functioning of AI. The chapter is a basic for my thesis for providing definitive and descriptive understanding of AI which are also crucial to comprehend discussions for functioning and behavior of AI and knowing how decisions are taken by AI. So, chapter will be an introduction for my hypothesis by showing that AI is a statistical machine that analyzes data given to it which is an important point for understanding bias and discrimination of AI. The issue in my argument specifically based on the "statistical machine" which cannot produce bias or discrimination intentionally and consciously.

## 2.1 BACKGROUND OF ARTIFICIAL INTELLIGENCE

To understand what AI is, it is essential to look at historical discussions. When the background of AI is considered, one of the most important names is Alan Turing. In his 1937 treatment, he used the specific definition of a computer with the concept of "a machine doing the same work as a human" (Larson 2021, 10). In the same period, Gödel has examined the possibility of "reducing human thinking to computation" (Larson 2021, 12). With Gödel's argument, Turing expanded his concepts, reversed notions of Gödel's. After 13 years from his first implication of AI, he proposed the Turing Test(Larson 2021,

18). In his 1950 paper, *Computing Machinery and Intelligence*, he provided a principal contribution where he discussed the thinking machine in a more precise framework. Turing proposed an altered version of a party game called the "Imitation Game" operated to answer the following question. This game includes three players; player A, player B, and player C. Throughout the game, player C—the interrogator standing apart from the other two—is tasked with determining which of the remaining two players is, player A or player B. Answers to questions with only one clue, man and a machine (Harnad 2008). A machine capable of misleading a person in thinking it was the human in the game would pass the "Turing Test". Although the original Turing Test is changed, the idea endures: "any computer holding a sustained and convincing conversation with a person would be doing something that requires thinking" (Larson 2021, 10). So, to pass Turing Test, a machine must mislead people into considering that they are interacting with a human when they are corresponding with a machine. As if claiming Alan Turing's call for a machine that can learn from experience, so AI can do just that(Press 2017). From Turing to today, many scholars such as Jack Good, Nick Bostrom, John Von Neumann, and Kevin Kelly, Kurzweil, etc. continued to argue various hypotheses for AI, and it still goes on (Larson 2021, 33- 49; Frana and Klein 2021, xvii-xxvi). A brief chronology of AI is given in Table 2.1.

**Table 2. 1: A Brief Chronology of Artificial Intelligence**

| | |
|---|---|
| The 1940s-1950s | Publication of Warren McCulloch and Walter Pitts' paper on a computational theory of neural networks, named "A Logical Calculus of the Ideas of Immanent in Nervous Activity."<br><br>Publication of Cybernetics, or Control and Communication in the Animal and the Machine by mathematician Norbert Wiener. |

| | |
|---|---|
| | Von Neumann hypothesized a self-reproducing machine require, at least, eight elements, including a "stimulus organ," a "fusing organ" to unite parts, a "cutting organ" to separate connections, and a "muscle" for action. |
| The 1950s-1960s | The Turing Test, attributing intelligence to any machine capable of exhibiting intelligent behavior equivalent to that of a human, is described in Alan Turing's "Computing Machinery and Intelligence."<br>Mathematician John von Neumann publishes "General and Logical Theory of Automata," reducing the human brain and central nervous system to a computing machine.<br><br>Artificial intelligence research begins at Carnegie Tech under economist Herbert Simon and graduate student Allen Newell.<br><br>Mathematician John McCarthy coins the term "artificial intelligence" in a Rockefeller Foundation proposal for a Dartmouth University conference.<br><br>The Dartmouth Summer Research Project, the "Constitutional Convention of AI," brings together experts in cybernetics, automata, information theory, operations research, and game theory. |

| | |
|---|---|
| | John McCarthy at the Massachusetts Institute of Technology (MIT) specifies the high-level programming language LISP for AI research. |
| The 1960s-1970s | The Stanford Artificial Intelligence Laboratory (SAIL) is founded by John McCarthy.<br><br>ELIZA, the first program for natural language communication with a machine ("chatbot"), is programmed by Joseph Weizenbaum at MIT.<br><br>The First International Joint Conference on Artificial Intelligence (IJCAI) is held in Washington, DC. |
| The 1970s-1980s | Arthur Miller publishes The Assault on Privacy: Computers, Data Banks, and Dossiers, early work on the social impact of computers.<br><br>John Holland uses the term "genetic algorithm" to describe evolutionary strategies in natural and artificial systems.<br><br>EXPERT, a generalized knowledge representation scheme for creating expert systems, becomes operational at Rutgers University. |
| The 1980s-1990s | The First National Conference of the American Association of Artificial |

| | |
|---|---|
| | Intelligence (AAAI) is held at Stanford University.<br><br>Philosopher John Searle makes his Chinese Room argument that a computer's simulation of behavior does not in itself demonstrate understanding, intentionality, or consciousness.<br><br>Computer scientist Doug Lenat starts the Cyc project to build a massive commonsense knowledge base and artificial intelligence architecture at the Microelectronics and Computer Consortium (MCC) in Austin, TX.<br><br>Marvin Minsky publishes The Society of Mind, which describes the brain as a set of cooperating agents. |
| The 1990s-2000s | IBM's Deep Blue supercomputer defeats reigning chess champion, Garry Kasparov under regular tournament conditions.<br><br>Dragon Systems releases Naturally Speaking, their first commercial speech recognition software product. |
| The 2000s-2010s | Visage Corporation debuts the FaceFINDER automated face-recognition system at Super Bowl XXXV.<br><br>Netflix announced a 1-million-dollar prize to the first programming team that |

| | |
|---|---|
| | develops the best recommender system based on a dataset of the previous user ratings.<br><br>Google begins its Self-Driving Car Project (now called Waymo) in the San Francisco Bay Area under Sebastian Thrun.<br><br>Stanford University computer scientist Fei-Fei Li presents her work on ImageNet, a collection of millions of hand-annotated images for training AIs to visually recognize the presence or absence of objects. |
| The 2010s-2020s | IBM's natural language computing system Watson defeats past Jeopardy champions Ken Jennings and Brad Rutter.<br><br>Apple releases the mobile recommendation assistant Siri on the iPhone 4S.<br><br>The Human Brain Project of the European Union is launched to understand how the human brain works and emulate its computational capabilities.<br><br>Facebook releases DeepFace deep learning facial recognition technology on its social media platform. |

| | |
|---|---|
| | Microsoft's artificial intelligence chatbot Tay is released on Twitter, where users train it to make offensive and inappropriate tweets.<br><br>The European Union publishes its General Data Protection Regulation (GDPR) and "Ethics Guidelines for Trustworthy AI." |

(Frana and Klein 2021; Larson 2021)

## 2.2 DEFININING ARTIFICIAL INTELLIGENCE

The term AI is coined by John McCarthy in 1955 (Frana and Klein 2021, xviii). In his revised article, *What is Artificial Intelligence*, he defined it as "the science and engineering of making intelligent machines, especially intelligent computer programs" (McCarthy 2007, 1). From McCarthy's first definition to today, scholars make multiple definitions. To provide a general outlook and comprehend the basics, some definitions are crucial. The definition of AI in *Encyclopedia Britannica*, which is a common source of terms, "the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings" (Copeland B J 2020). While Nilsson describes it as "the intelligent behavior in artifacts," while "intelligent behavior" involves "perception, reasoning, learning, communicating, and acting in complex environments" (1998, 1), Baltzan focuses on the facilitation of disorganized strategic decision-making, mimicking human thought, and behavior to simulate human intelligence (Baltzan 2013). A specialist organization established by the European Commission described that:

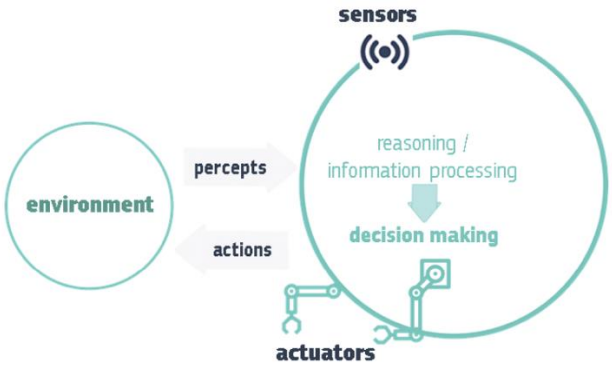
> "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or

> learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions.
>
> As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors, and actuators, as well as the integration of all other techniques into cyber-physical systems)(HLEG ECAI 2019, 6).

All these definitions show AI has a relation to mimic/reproduce[ii] cognition and creativity. There is a distinction between machine and natural intelligence. Herbert Simon set the boundaries for artificial intelligence by writing "...[A]rtificial things may imitate appearances in natural things while lacking, in one or many respects, the reality of the latter. Artificial things can be characterized in terms of functions, goals, adaptation…"(Simon 2019, 5). The relation between natural intelligence and artificial intelligence is inherently related to the idea of simulation but moreover "has focused on rational behavior [and thus] a machine is intelligent to the extent that what it does is likely to achieve what it wants, given what it has perceived" (Russell 2019, 41). As a field of computer science, AI is "concerned with designing intelligent computer systems, i.e., systems that exhibit the characteristics which we associate with intelligence in human behavior" (Barr and Feigenbaum 1981, 3). Otherwise stated, AI can be classified as the study of building machines and technology that can conduct activities that typically require human intelligence (Heath 2018).

After all these definitions and approaches, without neglecting the differences, there are common points in all of them. To put it briefly, the basics of what AI is in Fig. 2. 1:

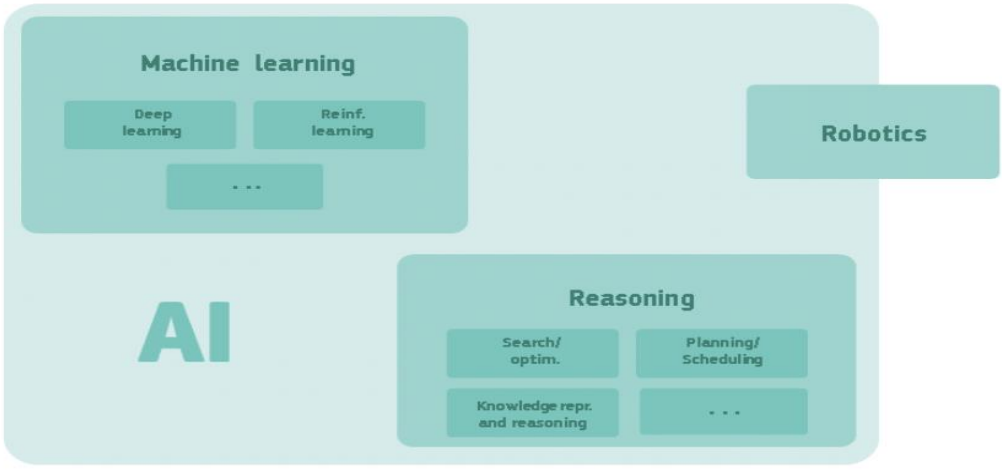
**Fig.2. 1: A schematic depiction of an AI system**




(HLEG ECAI 2019, 2)

So, artificial intelligence perceives the environment, which means dataset, and after processing the information, decides what to do by algorithms. AI relies on "algorithms that collect, analyze, de- and re-contextualize large data sets to explore and recognize patterns" (Strauß 2018, 3). AI has different types, such as soft computing, quantum computing, neural network, machine learning. (Chaturvedi 2008; Mermin 2007; Lu 2021). Even though categorizing and naming all these types shows differences from author to author, for categorization, I chose the approach of HLEG ACAI for the paper because it proposes proper and extensive outline.

**Fig.2. 2: Subsets of AI**




(HLEG ECAI 2019, 5)

According to Fig. 2. 2, there are three main subsets of AI. Robotics mainly refers to "AI in action physical world" (HLEG ECAI 2019, 3); reasoning is in relation to the modal logic and knowledge-based support decision system (Lu 2021). Machine learning (ML) is a term coined by Arthur Samuel in 1959. It means computer programs that can show behaviors more than programmed behavior (Joshi 2020). In other words, MLs displays 'learning,' which is related to intelligence.

## 2.3 MACHINE LEARNING

ML is one of the most progressive areas in AI. One reason behind the consideration of AI is the aim of designing an AI that "process potentially extensive and heterogeneous data

sets using complex methods modeled on human intelligence to arrive at a result which may be used in automated applications" (Datenethikkommission et al. 2018 as cited in Wischmeyer and Rademacher 2020, vii). ML, like human learning, is a process in which computers collect and process large amounts of data to identify patterns in the data (Emspak 2016). For that, learning of AI presents an indispensable role. ML picks up the models and imitates human intelligence as well as can update itself through analyses. Following repetitions and alteration of the algorithm, the machine input and predicts an output (Naylor 2018; Bini 2018). Thus, ML is not just related to the statistics tool, classification, or identification; ML means the ability to learn and simulate behavior.

To understand the ways of simulating behaviors, models of learning are essential. Three methods of ML learning, supervised learning, unsupervised learning, and reinforcement learning, respectively, are explained.

Supervised ML requires a data set with labeled samples and guidance. By recognizing labels and examples, or namely through classification, the machine learns to perform to the desired behavior. On the contrary to supervised ML, unsupervised ML does not include labels. Unsupervised ML tries to find patterns in the data (Joshi 2020). In reinforcement learning (RL), the machine performs actions in a defined environment and gets feedback to guide its behavior (Marwala 2021; Joshi 2020). Different methods for learning of AI are given in Table 2.2.

**Table 2. 2: Models of ML**

| **Models of Learning** | **Method Samples** |
|---|---|
| Supervised Learning | Case-based learning[iii]<br>Bayesian networks[iv]<br>Decision-tree inductions[v]<br>Linear regression[vi] |
| Unsupervised Learning | K-means clustering[vii]<br>Self-organizing maps[viii]<br>Neural networks[ix]<br>Genetic algorithm[x]<br>Deep learning[xi] |

| | |
|---|---|
| Reinforcement Learning | Value iteration[xii]<br>TD Learning[xiii]<br>Q Leaming[xiv] |

Different learning methods of ML and applications of these methods result in the use of AI in different areas. Nowadays, machine learning applications can be used for many areas such as health, natural language processing[xv] (NLP), security, insurance, human resources, law,. AI's results are affected by models, that is why choosing the appropriate model for the desired performance is a top priority.

When one understands what AI is and how it learns, the most extreme problem of AI can be seen as simulating human actions. However, even imitation of AI depends on what it takes from the data and learns from the model. It easily shows that AI's bias cannot depend on AI's operation or processing. AI decides based on the model that it learns to use rules, tags, patterns but none of these methods is data/training model independent. From this point of view, it is necessary to clarify how bias emerges and how discriminatory decisions it make, by accepting that AI cannot engage in discriminatory behavior or be "self-biased".

# 3. FRAMING DISCRIMINATION AND BIAS FOR ARTIFICIAL INTELLIGENCE

In the previous chapter, by using literature, I defined what is AI and how AI learns. Historical development of AI both academically and practically is given in Chapter 2. So, AI is an agent that takes a role in the decision-making process. By description, they are not human. As artificial agents, they do not carry elements that are needed for morality. Righteousness, virtues, harm, or emotions do not influence their preferences. However, the human brain cannot process all the data to compose a rational choice and is riddled with biases (Kahneman 2011).

AI can receive an immense quantity of data for a rational decision. From this perspective, intelligence that is free from moral elements and bounded to the mathematical inputs cannot manage biased or discriminatory activities, so to speak, the decisions of AI are rational. However, recent research has shown that in many AI-studied areas, decisions made by AI discriminate against people and marginalize people. While the decisions made by AI can be discriminatory, it is also important to examine AI's intentionality, deliberation, and consciousness. In this section, literature on discrimination and bias is presented. Definitions of bias and discrimination, and different approaches to the terms, are stated. Different bias types are also defined in the chapter. Understanding the perspective of literature for both terms and examining bias types are crucial for the structure of thesis. This chapter provides the basic information and theoretical background for my hypothesis through the literature.

## 3.1 DISCRIMINATION OF AI

In the *Weapons of Math Destruction* written by Cathy O'Neil (2016), the main question can be summarized as do people over trust AI? Throughout the mortgage crisis, minority groups such as blacks or the poor were victims of AI. Much software system models, which frequently manage business environments, are marked with human prejudice,

dislikes, and bias. O'Neil declares multiple representations of discriminatory actions of AI (O'Neil 2016). In AI activities, biases against women, people of color, poor people, and LGBTQ people are found. These activities include facial recognition, Fintech loan approval determination, anti-discrimination social media tools, and search results for job opportunities (Dalenberg 2018; Howard and Borenstein 2018; Wilner 2018). Racial and gender bias is crucial because these discriminations are systemic in diverse aspects of society (Ennals 2016).

Representatives of bias in artificial intelligence can be observed in applications extending from beauty contests deciding to sentence algorithms. Empirical evidence has explicated AI is biased against preserved classes (Lee 2018), while another study proved that the predictive algorithms are fallacious with societal stereotypes, by pointing African–Americans are more likely to act violent crimes than whites (Kirchner et al. 2016). The results of the research conducted by Buolamwini and Gebru (2018) show that facial recognition software embedded in smartphones works best for white-male people. Leavy (2018) revealed how AI learns bias from the text. Gender bias is represented in word embedding (Bolukbasi et al. 2016). Also, with word embedding, racial and gender bias has been showed (Caliskan, Bryson, and Narayanan 2017). The recruiting algorithm focused on male candidates at Amazon, which is learned from Amazon's male-tended preferences in human resources (Dastin 2018). For example, by training Google's algorithms on historical human-generated search queries, the algorithms were taught to make false prejudiced and abusive colorations about people or groups in both search outcomes and the targeted ads (Osoba and Welser IV 2017). In an obvious way, many studies provide shreds of evidence on the discriminatory behaviors and prejudices of AI within several domains.The consequences of artificial intelligence decisions are predominantly racist, sexist, or age-based attitudes.

### 3. 2 BIAS TYPES

In AI and ML, bias refers to prior information, which is necessary for intelligence (Bishop 2006). AI's bias and reflections can be seen in many fields, such as the justice system, search engine application, voice recognition, and robotics (Howard and Borenstein 2018).

Although many scholars study bias, one problem with AI bias is that there is no consensus in the literature. For that reason, different approaches and definitions of bias is given in Table 3.1.

**Table 3.1: Types of Bias**

| Types of Bias | Definition |
|---|---|
| Absolute Bias | "is an assumption by the learning algorithm that the target function to be learned is definitely a member of some designated set of functions" (Dietterich and Kong 1995, 1). |
| Aggregation Bias | arises during model construction, when different communities or groups are inappropriately combined (Suresh and Guttag 2019)[xvi]. |
| Algorithmic Bias | is the behavior in autonomous systems or machine learning-based practices, which harms users based on gender, race, or disability. It can enter and happen when the bias is not present in the input data, or modeling, testing, etc. Then the obtained bias is processed and calculated by the algorithm (Aysolmaz, Iren, and Dau 2020; Baeza-Yates 2018; Danks and London 2017; Lee, Resnick, and Barton 2019; Olteanu et al. 2019). |
| Behavioral Bias | is methodical falsifications in user behavior across platforms or contexts, or across users depicted in different datasets (Olteanu et al. 2019). Content production bias[xvii] and linking bias[xviii] are subsets of behavioral bias (Mehrabi et al. 2019, 4). Filter bubbles and personalization are samples for behavioral bias (Bozdag 2013). |
| Correlation Bias | happens when irregular or unpremeditated connections emerge through data processing. It is also related to reflect societal bias and incorporates with it (Gregorutti, Michel, and Saint-Pierre 2017; Woosley and Sherman 2019). |
| Cause-Effect Bias | arise as a consequence of the fallacy that correlation signifies causation (Mester 2017). |

| | |
|---|---|
| Content Production Bias | behavioral biases are expressed as lexical, syntactic, semantic, and structural differences in the content generated by users (Olteanu et al. 2019). |
| Data Bias | lack in the Big Data's 5V[xix] can cause false correlations of data or misrepresentations and it named as data bias (Barocas and Selbst 2016). |
| Emergent Bias | appears as an outcome of the use and trust of algorithms across new or unanticipated circumstances. The shift in population, cultural preferences, or societal experience or knowledge which is happened after the end of design can conclude as emergent bias (Friedman, Kahn, and Borning 2008). |
| Evaluation Bias | arises when the testing or outside benchmark populations do not fairly portray the multiple parts of the user population (Suresh and Guttag 2019). |
| Explicit Bias | "Biases resulting from factors outside the social platform, including considerations of socioeconomic status, ideological/religious/political leaning, education, personality, culture, social pressure, privacy concerns, and external events" (Olteanu et al. 2019)[xx]. |
| Historical Bias | already existing bias in the world is propagated in a model (Suresh and Guttag 2019). |
| Implicit Bias | is stances or conventions that influence perception, behaviors, and decisions unconsciously (Lee 2018). By using word-embedding, bias happened because of the gendered semantics and reflected in AI is an example of implicit bias (Caliskan, Bryson, and Narayanan 2017). |
| Interaction Bias | grows in AI from the users' interactions and selections. Chatbot that was shut down because of using racial slurs is an example of interaction bias (Woosley and Sherman 2019). |

| | |
|---|---|
| Measurement Bias | appears when preferring and measuring features and labels to use; these are often proxies for the desired quantities (Suresh and Guttag 2019). |
| Omitted Variable Bias | occurs when one or more important variables are left out of the model (Mester 2017). |
| Popularity Bias | occurs in algorithms prioritizing popularity metrics to rate or suggest content to users. More popular items tend to be presented more (Ciampaglia et al. 2018). |
| Population Bias | systematic distortions in demographics or other user characteristics between a population of users represented in a dataset or on a platform and some target population (Olteanu et al. 2019). |
| Pre-existing Bias | is a bias enduring in society, which is carried into the algorithm (Howard and Borenstein 2018). |
| Ranking Bias | is related to the idea that the top-ranked results are most relevant and essential. As a result, the content of interest will show more frequently and will receive more clicks than others (Baeza-Yates 2018). |
| Relative Bias | "is an assumption that the function to learned is more likely to be from one set of functions than from another" (Dietterich and Kong 1995, 1). |
| Representation Bias | appears in "defining and sampling a development population" (Cobb and Bock 1994; Suresh and Guttag 2019). |
| Sampling Bias | appears due to the non-random sampling of subgroups and mistakes in the choice of data (Mehrabi et al. 2019). |
| Technical Bias | develops through the constraints of a program, compute power, design, or other limitations in the system(Bozdag 2013). |
| Temporal Bias | is systematic distortions across user populations or behaviors over time (Olteanu et al. 2019). |

| | |
|---|---|
| User Interaction Bias | is a type that is impacted by two sources: interface and through the user itself by imposing his/her self-selected bias (Baeza-Yates 2018). |

It is easy to see in the table that there are various approaches and separate definitions for AI bias. Though, the relationships and similarities between kinds are also considerably noticeable. Therefore, in this article, a more precise and definite categorization is given to demonstrate the relationship and differences of bias types. By considering the similarities, the operation process of artificial intelligence, and the differences, three main categories are determined, and the previously identified bias types are placed under the appropriate categories. The recategorization of the definitions of bias given above is a requirement for a simpler and more understandable ground. This classification will also serve as a basis and guide to discuss the source of the bias. When these definitions, which are scattered but related to each other, are combined in a more regular categorization, their relations with each other and the connection of many different seeming bias types with society, data and technical infrastructure will be revealed. By helping to establish a causal relationship with the source of bias, recategorizing presents a concrete structure for discussing one of the main points of the argument, the connection of bias to the socioeconomic system. Each category will be named and explained in the section below.

### 3.2.1 Inherited bias

This category refers to types of bias that are identified with an obvious association with society. Historical bias, explicit bias, implicit bias, and pre-existing bias; belong to inherited bias. None of the relevant types of bias arise concerning the operational process of artificial intelligence, and they are all representations of what already exists in society. Some researchers have found that societal stereotypes, such as African lead numerous predictive algorithms–Americans are keen to commit violent crimes (Kirchner et al. 2016). The algorithm's evolution or re-training from its primary code corresponds with power structures, societal expectations, assumptions, and preferences (Noble 2018). The associations made by humans to decide or problem-solving cause biases implicitly or explicitly. These biases have been institutionalized in society and used to perform group

classifications and discrimination toward out-group members (Daumeyer et al. 2019; Katyal 2019; Barocas and Selbst 2016). Algorithms acquire data from the past and are trained on societal knowledge, which concludes as a reproduction of bias (Barocas and Selbst 2016; Mayson 2018). Additionally, research also revealed that algorithms reconstruct the biases of their creators (Howard and Borenstein 2018). To conclude, types that are named under inherited bias are emerged in society, not in the model.

### 3.2.2 Dataset bias

This category defines the kinds of bias that are bound to the data generation and construction process. Behavioral bias, correlation bias, cause-effect bias, content production bias, data bias, evaluation bias, interaction bias, measurement bias, omitted variable bias, population bias, ranking bias, representation bias, sampling bias, and user interaction bias fall into this section. It is valuable to examine dataset bias under two attributes. First, data bias, population bias, representation bias is directly operative in the data production process, influencing the quality and heterogeneity of the data. On the other hand, behavioral bias, correlation bias, measurement bias arises from the interrelationship of data rather than from the production of data; that is, they are the consequences of the data production process. Secondly, several types are more interested in the diversity and representativeness of the data. While the data is not affected by the possible variables, it depends on the social acceptances and judgments before the data acquisition process. On the other hand, other types are instantly influenced by the users' approaches, prejudices, evaluations, and actions.

At this point, it is essential to examine big data for a more grounded discussion on dataset bias. Big data, partial data, or overrepresentation in data can lead to ‘disparate’ treatment or unjustified bias of protected classes (Barocas and Selbst 2016; Lee 2018). One of the leading causes for dataset bias is the nonrandom, systemic neglect of people who live on big data's edges, whether due to poverty, geography, or lifestyle, and whose lives are less 'datafied' than the overall population's (Lerman 2013). To exemplify, in 2016, Airbnb published that there are hosts denied because of race, age, gender, and additional representatives (Murphy 2016). The bias in the dataset is an artifact of the data mining

process itself (Barocas and Selbst 2016). Non-representative samples in the data or standard validation of data for all groups can result in systematic errors and bias (Barocas et al. 2017). In other words, dataset bias is assessed by what the data includes or excludes and how well it corresponds to big data standards.

### 3.2.3 Technical bias

This category defines the types of bias correlated with modeling. Restraints that affect outcomes in algorithms, model building, and mathematical equations fall into this category. Aggregation bias, absolute bias, algorithmic bias, emergent bias, popularity bias, relative bias, and temporal bias fall under technical bias. The main issue with the technical bias category, it is highly possible that replicate structural or explicit bias or generate new ones (Lee 2018). "This form of bias originates from all the tools used in the process of turning data into a model that can make predictions"(Dobbe et al. 2018, 2).

In chapter 3, by using literature, I defined discrimination and bias. After stating types of bias, I categorized bias types.

# 4. PROPOSING MODELS

In Chapter 3, I put the main definitions and approaches to bias and discrimination. However, the complexity of bias types and the lack of relationship between bias and discrimination of AI are major gaps in the literature. In this chapter, I propose two distinct but related models to close the gap. I provide a framework for the relationship between discrimination and bias in AI in the first model. I Re-categorized bias types with the perspective of social shaping of technology in the second model. Models crucial to understanding how AI bias and society's judgments are related to each other. The chapter underpins the point in the argument of my thesis that the bias of society is inherent to data and transmitted to AI, rather than directly the problem of AI functioning or logic. My models depend on the perspective of social shaping of technology theory. Without the perspective of SST, the logic of models cannot be understood clearly. So, relation between society, bias and discrimination is clarified in this chapter.

## 4.1 An Approach to Technology and Society

*"Every piece of information you obtain on one system is also information on other" (Latour 1987, 138)*

It is essential to develop an approach to technology to explain how bias is related to society. For instrumental theory, technologies are “tools” that stand ready to follow the objects of their users. Technology is supposed "neutral" with no valuable content of its own. From this point of view, technology is independent of politics, economic structure, or society. Another implication of this position is the "neutrality and rationality of technology", meaning that technological tools "maintain their cognitive status in every conceivable situation" and "mainly under the same norm of efficiency in all contexts". Also, "Its universality, therefore, means that the same measurement standards can be applied to it in different environments” (Feenberg 2002, 5-6). A different theory, technological determinism, is similarly important in new media. Technological determinism is the concept that technological development is autonomous from society;

it shapes society, although the influence is not two-sided. Technology endures outside of society, yet it affects the social environment. Technological determinism puts the existence of technologies as the final determinant at the core of changes in social formations and action orientations. In its most uncomplicated form, it describes the changes that occur in human societies and cultures because of technological development or differentiation (Baştan 2017). According to Bell (1972, 92), the determining power behind contemporary forms of social organization is new intellectual technologies. New intellectual technologies are changing our modes of experience, interactions, identities, and time orientations, leading to fragmentation both between social structures and culture and between cognitive and emotional expressions. In more severe forms of technological determinism, technology is perceived as the most important determinant of the nature of society (Mackay and Gillespie 1992).

For a long time, deterministic approaches have dominated discussions of society and technology. However, SST focuses on social change, roles of meanings, knowledge and power relationship, technology and knowledge hierarchy. SST has thus moved away from the concept of 'effects', which are simply outcomes determined by the character of technology. Technology can maintain an analysis of the interests represented in technology. Among the other influences that shape society with technology, there are general discourses about technology, especially deterministic and utopian discourses that emphasize the neutrality, inevitability, or rationality of technological change.

So, the social shaping of technology theory explains that the development of a particular technology needs a comprehensible model of the society in which the technology is embedded. The very construction and structure of current information technology is itself a product of historical, social, and economic shaping (Williams and Edge 1996). Williams says that technological development depends on an interaction that is included in socio-cultural structures framed by commercial, military and political purposes. Technology can achieve an effective position only when it is used for purposes that a known social process already contains. Theory denotes that technology can" embody specific forms of power and authority" (Winner 1980, 121).

Addressing technology within concepts of efficiency and productivity only is not reliable. It is crucial to debate the circumstances that are 'internal' to the operations of a given technical system and 'external' to it (Winner 1980, 130). In brief, technology is constantly connected to forms of authority and exhibits power structures in society. Alternatively stated:

> "... close inspection of technological development reveals that technology leads a double life, one which conforms to the intentions of designers and interests of power and another which contradicts them- proceeding behind the backs of their architects to yield unintended consequences and unanticipated possibilities". (Noble 2017)

Distinct peculiarities in the technological design or its components can present valuable means of organizing patterns of power and authority in a presented setting (Winner 1980). Each step in the production and implementation of new technologies involves various options among different technical options. Besides narrow 'technical' considerations, some 'social factors' force which selections are preferred, affecting the content and social influence of technologies. Social shaping theory explores the social settings of innovation (from design to development of technology) social and economic forces that can form technology (a division of labor, and expertise within and between organizational structures; industry and market structures; etc.) and the position of an extensive spectrum of associated and interested groups (including not only technologists and decision-makers yet further end-users) (Williams and Edge 1996). Technology, which is scientific output and human product, cannot be thought of without the knowledge of society and the prerequisite of technological development in society.

> "One can see the empirical world only through some scheme or image of it. The entire act of scientific study is oriented and shaped by the underlying picture of the empirical world that is used. This picture sets the selection and formulation of problems, the determination of what are data, the means to be used in getting data, the kinds of relations sought between data, and the forms in which propositions are cast…[t]he underlying picture of the empirical world is always capable of identification in the form of a set of premises. These premises are constituted by the nature given either explicitly or implicitly to the key objects that comprise the picture". (Blumer 1986, 24-25)

The approach that technology is socially shaped is radically opposed to technological deterministic explanations of the nature of technology, the relationship between a society and its technologies, and surely the foundations of the social system and the origins of social change.

One of the most important debates on technological determinism is the idea of rationality. In terms of technological determinism, AI can be viewed as “a rational machine is a device designed to maximize its performance to achieve its goal”(Marwala 2021, 31). Detailly explained:

> “Rational decision-making is a process of making decisions using logic and reason to maximize the net utility. Decision-making is a process of attaining a decision and comprises many possible decision outcomes that is rational if it maximizes the net utility”(Marwala 2021, 47).

While this concept of rationality is logical, it ignores the sociological aspect of rationality. To discuss rationality, it is important to look at Weber. Weber distinguished two different kinds of rationality, “corresponding to social thought and action. Rationality is 'substantial' to the extent that it fulfills a particular value” such as sustaining social authority. “The ‘formal’ rationality of capitalism introduces economic arrangements that optimize calculability and control”. However, Weber’s approach is criticized by Marcuse, and he proposed ‘technological rationality’ which constitutes the basis for elite control of society. That control is not simply an extrinsic purpose served by neutral systems and machines but is internal to their very structure” (Feenberg 2002, 65 - 66).

The perspective and conceptualization presented by Marcuse are important because it objects from a point of view that is included in formal rationality and points to its structural accumulation. Just as it is necessary not to evaluate technology and social shaping only as a superficial association, technological rationality also considers technology as a set of meanings and values that have played a role in the formation of technology. It is also important to look at the information neutrality discussion to make the perspective more grounded.

Technological determinism also refers “data/knowledge neutrality”. As Foucault puts, stated in Feenberg “knowledge and technology are not value-free tools that may be put to a good or bad use. Truth and power are not two independent things that meet contingently in the moment of application”(2002, 68). Additionally, another point of Foucault “It is the actual instruments that form and accumulate knowledge, the observational methods, the recording techniques, the investigative research procedures, the verification mechanisms. That is, the delicate mechanisms of power cannot function unless knowledge, or rather knowledge apparatuses, are formed, organized, and put into

circulation" (Foucault and Ewald 2003, 34) Haraway also questioned the power relations with the context of "situated knowledge" by stating

> "How to see? Where to see from? What limits to vision? What to see for? Whom to see with? Who gets to have more than one point of view? Who gets blinded? Who wears blinders? Who interprets the visual field? What other sensory powers do we wish to cultivate besides vision?"(Haraway 1988, 587).

Knowledge production, politics in knowledge, knowing subjects are all questioning of Haraway. Ultimate rejection the technical reconstruction of the entire field of social relations within which it operates.

> "The power of the businessman or bureaucrat is already present in the fragmentation of the various social spheres of production, management and labor, family and home life, economics and politics, and so on. The fragmented individuals and institutions can be organized only by agents who dominate them from above"(Feenberg 2002, 183).

By seeing all these rejections, the bond between structure of society and technology becomes clear. To put it other words, technological designs and technological content are also social designs and contextualized by society. Cultural values, economic interests, and political decisions are as integral to their composition as mathematical calculations, and technologies are extensions of structures of power and capital, and derivatives of scientific and engineering discourses.

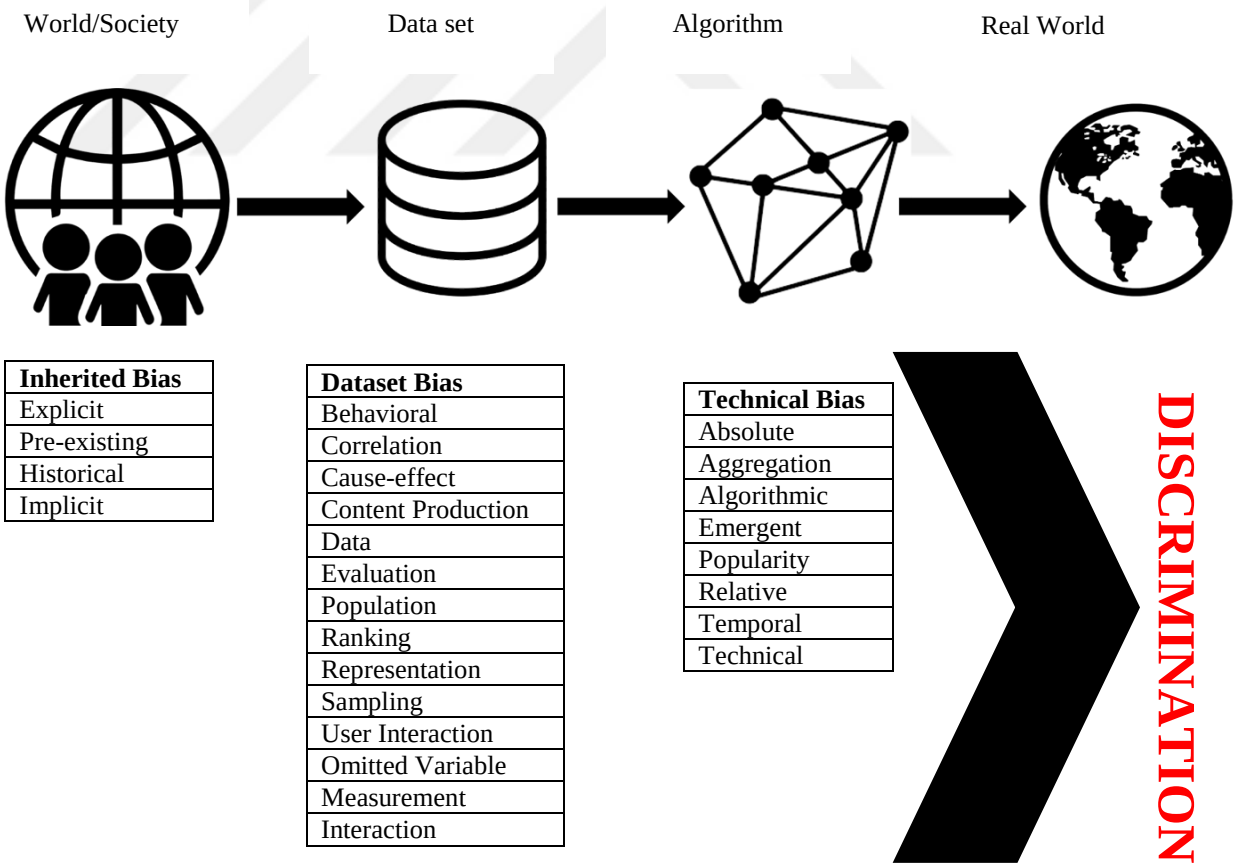
Therefore, evaluating it in terms of social shaping theory and considering its methodological approaches to society and technology relationship provides a way to understand AI bias. AI's bias is significantly produced by the public, and attitudes performed by society are coded into models by data and sometimes by the influence of those in charge. As my argument suggests, bias is contained by data and transmitted through data to AI, and bias of AI is related with societal structure, it is shaped by social settings and engagement of technology with economic, sociological, cultural is a major determinant.

Social shaping of technology provides a framework to understand the relationship between technology and society. With the perspective of SST, "bias of AI" becomes a societal problem, rather than technical issue.

### 4.2 Two Models for Bias and Discrimination

By considering the terms of bias, it is critical to comprehend its associations with the discriminatory behaviors of AI. The bias of AI is rooted in societal prejudices, dataset inclusions, and exclusions, or technical reasons. Still, the bias of AI is a notion within AI, not out. The outcome of bias is outside of it. Therefore, the "bias inside AI learning" is reflected in "discrimination" in real-world results. The model of the discussion can be seen in Fig.4.1.

**Fig. 4.1: FBTD (From Bias to Discrimination)**



The logic of my FTBD model is that discrimination cannot be thought of as something AI does. As a statistical machine, AI cannot act deliberately or with the intention of discrimination. The reason I put forward this model is to represent the relationship between bias and discrimination. When biases arising from society, dataset, or technical infrastructure, which I categorized in the previous section, are processed by artificial intelligence, the actions that are called socially discriminatory are essentially mathematical decisions made by artificial intelligence in line with the inputs. Discrimination, which is a sociological definition, emerges in the real world on the axis of the relationship of artificial intelligence outputs with the social values that do not exist in the statistics of artificial intelligence. In other words, the decisions made when the statistical result of artificial intelligence meets the values of the real world are discriminatory, but this discrimination is not inherent in the logic and functioning of artificial intelligence. On the other hand, different types of biases in artificial intelligence affect the formation of these results. Many biases in the data and modeling of artificial intelligence affect the results of artificial intelligence. However, on the other hand, it is not possible to attribute the discriminatory quality to the bias in the data or modeling of artificial intelligence until the artificial intelligence gives an output.

In this context, FBTD reveals that discrimination is not a behavior of artificial intelligence, but a quality that emerges in real-world results; however, it shows that the functioning of different types of bias in artificial intelligence is also the factor that causes these results. My model is formed to create a new discussion ground for many approaches in the literature states that artificial intelligence has discriminatory behaviors. With more accurate concepts and a cleaner ground, the model simplifies the conceptual complexity of understanding the source of the bias and discrimination in AI results.

The systematic presentation in FBTD combines the stages of the AI process and how AI bias manifests in different steps. Considered through the stages of appearance, Fig. 3 is inclusive and informative, although it hides a critical issue: the relationship between inherited bias and dataset bias. The question is, is dataset bias a subset of inherited bias? In the literature, dataset bias (including subtypes) is mainly examined in terms of

representation. However, even if the dataset comprises sufficient representation, will it be unbiased?

Artificial intelligence learns through data based on human history, Internet content, coders' assumptions, and model builders. Prejudices, socially discriminatory behaviors, or people's acceptance all flow into the data and show up as data flaws. Consequently, the data are not objective merely indicate existing social and cultural biases. Discrimination reflected in real-world results cannot be evaluated and considered by overlooking the relationship with the people from whom the data is collected. Data may contain bias through language, actions, or labeling (Caliskan, Bryson, and Narayanan 2017; Barocas and Selbst 2016). AI processes data from humans and model trains through data that is already biased. Thus, 'inherited bias independent' dataset bias is an illusion that ignores the authenticity of the data. Data circumlocutorily acquires societal bias and transfers it into the AI learning process. All discriminatory actions and behaviors in the data processed in AI are simulated, and societal biases are reflected in decisions. A new model focuses on the relations of bias types within this perspective in RBTD model.

**Fig.4.2: RBTD ( Re-categorized Bias to Discrimination)**

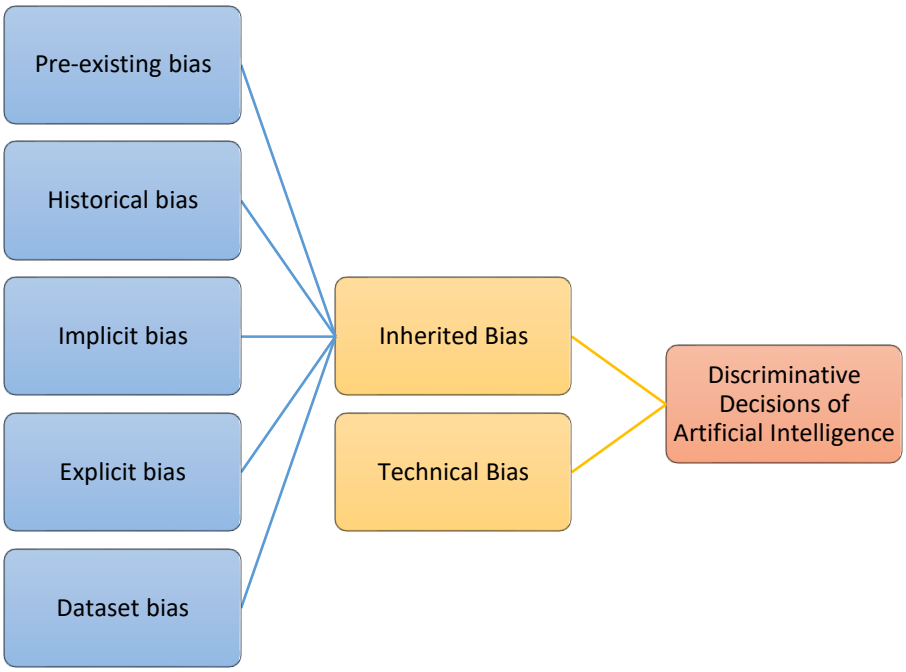


Fig. 4.2 presents a model for sources of bias based on machine and society. The model operates as a vital position in terms of technological determinism and the social shaping of technology. Without attempting a comprehensive design for AI bias, admitting bias as an issue associated with data undermines the social inferences. RBTD especially states that dataset bias must be categorized in inherited bias and shows that the sources of bias must be classified with two classifications. It is obvious that an approach that ignores the connection of society and dataset bias, which was previously referred to as a category, is insufficient in producing solutions and developing an approach to bias in artificial intelligence. Presenting the "dataset" bias as a particular category from a position that ignores discussions such as the hierarchy of knowledge, gendered information, the relationship between knowledge and power, and data neutrality aims to reduce the bias of artificial intelligence to a technical discussion. For this reason, my RBTD model makes sense of the relationship of data with various sociological concepts and to reveal its connection with society. Theoretical discussions on data will be further elaborated below yet the model reveals that artificial intelligence bias must be examined in two categories, by adding data set bias to the bias types that were under the inherited bias category in the previous categorization. Additionally, RBTD also refers to FBDT, showing that the real-world counterpart of the consequences of artificial intelligence in the discussion about artificial intelligence discriminating may be discrimination and the simple operation of discrimination in the literature. In summary, while the subtypes of inherited bias in the previous categorization were taken into consideration together with its social ties in the literature, in the model I developed, dataset bias was placed as a subtype of inherited bias based on its ties to the society, therefore, it was stated that there are two types of bias in artificial intelligence as inherited bias and technical bias. Finally, it has been shown that the quality of discrimination occurs with real-world values in real-world outcomes. These two models presented above, are simple grounds to show that there is no inherent bias in artificial intelligence and that bias and discrimination are community-based and community-based.

# 5. A CLOSER LOOK TO ARTIFICIAL INTELLIGENCE ECOSYSTEM IN TURKEY

In this part of the thesis, the details of the research are given. First, the importance of the Turkish ecosystem is mentioned and TRAI is explained. Afterwards, information about the interviewees is given. Answers of interviewers are stated and results are explained.

## 5.1 AN INTRODUCTION ECOSYSTEM OF TURKEY

In 2019, the findings of research conducted by McKinsey Global (Cam, Chui, and Hall 2019) showed a year-over-year increase of approximately 25 percent in the use of AI in standard business processes, and a significant increase in companies using AI in multiple areas of their business compared to last year. In the business areas where it is used, 44 percent state that it provides an increase in revenue and that artificial intelligence reduces costs.

The results also reveal that a small fraction of companies from various industries are achieving large-scale business results from AI. For AI sector of Turkey, results of the "Artificial Intelligence in Middle East and Africa" report (‘Yapay Zeka Kullanımı Raporu’ n.d.), which was prepared by interviews with the executives of more than 100 companies operating in 5 countries including Turkey (Turkey, Saudi Arabia, United Arab Emirates, Jordan and South Africa) reveals important points.

According to the report prepared in cooperation with Microsoft and EY to evaluate the use of artificial intelligence by companies in the Middle East and Africa region, Turkey, which is the leader in its region, became the country that invested the most in artificial intelligence in the region in terms of strategic importance, areas of use, awareness, and investments.

It was stated that 80 percent of companies in Turkey handle their artificial intelligence strategies directly in senior management. While 25% of companies consider AI among their strategic digital priorities, 60% recognize the importance of AI for their core

business. 35% of companies in Turkey actively use pilot artificial intelligence technologies compared to 28% in the region.

Companies benefit most from machine learning, according to usage intensity. While the rate of benefiting from machine learning is 61% in the Middle East and Africa region, it is around 85% in Turkey. 80% of companies in Turkey that integrate artificial intelligence into their operations expect effective benefits. At the beginning of the benefits of artificial intelligence; optimization of operations, digital transformation of products and services, empowerment of employees and being closer to customers are coming.

For the sake of study, the ecosystem of Turkey Artificial Intelligence Initiative (TRAI) is used. (TRAI) is an initiative established in 2017 to increase awareness of artificial intelligence and develop the ecosystem in Turkey. TRAI helps start-ups meet with academia and the private sector, make joint projects, find investors and expand abroad. In addition, TRAI has a Technology Advisory Board, a Sectoral Advisory Board, an Academic Advisory Board, and an Investor Committee. TRAI also conducts focused and concentrated studies with many working groups such as production & AI, health & AI, security & AI. Four objectives have been determined by TRAI. These goals are awareness, capacity, commercialization, and ethics, respectively. Awareness aims to contribute to the formation of a conscious society that is aware of the importance of artificial intelligence, its opportunities and threats, and its application potential in all fields. Another objective, capacity tries to contribute to the formation of competent people and institutions working in the field of artificial intelligence; encourage collaborations; increase the sharing of knowledge and experience. Commercialization is specifically related to employment, added value, and sustainability. Ethics aims to identify possible risks and threats with the development of artificial intelligence, raise awareness on this issue, and contribute to the elimination of risks. (‘Hakkımızda - Türkiye Yapay Zekâ Inisiyatifi’ n.d.)

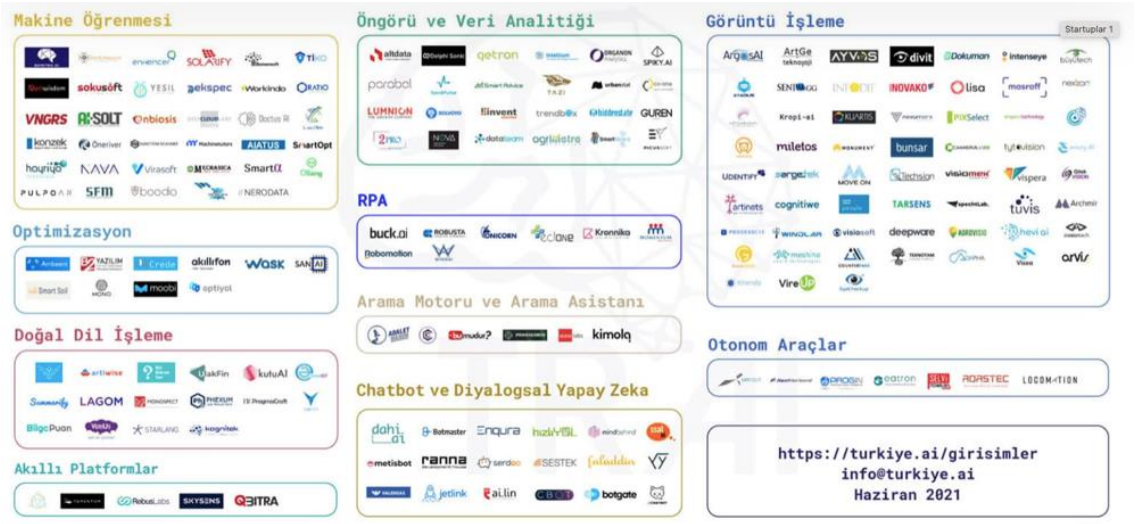
In the TRAI ecosystem, there are more than 50 private sector companies as well as startups. Close studies are carried out with the teams of these companies, contributing to the development and spread of artificial intelligence in our country. Artificial intelligence

summits and Turkey Artificial Intelligence Week events are held by TRAI (‘Etkinlikler - Türkiye Yapay Zekâ Inisiyatifi’ n.d.) .

The TRAI ecosystem consists of supporters, start-ups, scale-ups, technology partners, academic partners, TRAI fellows, and TRAI Community. Technology business partners include companies such as Intel, Amazon, and Google that gives infrastructure support to startups. Academic partners are open to faculty members working in artificial intelligence at universities. TRAI Fellows are open to professionals and experts who contribute to AI studies, while TRAI Community is open to university societies and NGOs producing content on technology(‘Hakkımızda - Türkiye Yapay Zekâ Inisiyatifi’ n.d.).

When we look at start-ups, there are over 100 startups, although the number is constantly changing. These startups are subject to 10 different classifications by TRAI according to the field of operation of the companies. Although some companies named under more than one classification, the relevant categories are: Machine Learning, Prediction and Data Analytics, Image Processing, Optimization, RPA, Natural Language Processing, Search Engine and Search Assistant, Autonomous Tools, Intelligent Platforms, and Chatbot and Dialogic Artificial Intelligence (‘Startuplar - Türkiye Yapay Zekâ Inisiyatifi’ n.d.). The Fig. 4.1 presents the companies in TRAI.

**Fig.5.1: Ecosystem of TRAI**




(‘Startuplar - Türkiye Yapay Zekâ Inisiyatifi’ n.d.)

When we look at the company distributions in the relevant categories, more than half of the companies are in the Image Processing category. After the Image Processing category, most of the companies included in the Machine Learning category. Foresight and Data Analytics and Natural Language Processing categories also have a certain majority among companies, while 18 companies appear in Chatbot and Dialogic AI. The number of companies operating in the categories of Optimization, RPA, Autonomous Vehicles, Search Engine, and Search Assistant and Smart Platforms is below 10. For the research, people from companies in the categories of ML, Foresight and Data Analytics, NLP, Chatbot and Dialogue Intelligence, RPA, Image Processing, Search Engine, and Search Assistant are interviewed.

## 5.3 A CLOSER LOOK TO INTERVIWEES

Seven people were interviewed for the study. Although the job descriptions of all the interviewees are different, the focus studies of the companies are also distinct from each other. To appreciate the scope of the research, it is imperative to acknowledge interviewers' jobs and businesses sincerely. Furthermore, the ecosystem problems are presented within the interviewers' words, giving a framework and suggesting implications to comprehend their discussion on the bias.

**Fig.5.2: Information Table About Informants**

| Interviewees | Age | Job | Position | Years (In field) | Years (In company) | Field of company | Use of AI | Market of Company |
|---|---|---|---|---|---|---|---|---|
| | 35+ | Computer Engineering | Chief Analyst | 13 | 4 | Advanced Solution | ML<br>Insight and Data Analytics | Global |
| | 30+ | Industrial Engineer | Sales and Marketing Director | 3 | 3 | Automation | NLP<br>Chatbot<br>RPA<br>ML | Local |
| | 30+ | Control and Automation Engineering | Product Manager | 4 | 1 | Retail | Image processing | Global |
| | 35+ | Academician | President | 9 | 2 | Consulting Company | Image processing<br>Search Engine | Estonia based company |
| | 50+ | Electronics Engineering | General Manager | 20 | 2 | Content Management | Neural networks<br>ML<br>NLP,ML | Local |
| | 25+ | Computer Engineering | AI Researcher | 2 | 2 | Security company | Image processing<br>Image processing | Global |
| | 25+ | Software Engineering | Machine Learning Engineer | 4 | 2 | Security company | ML | Global |

Information on age, occupation, position, duration of work on artificial intelligence, working time in the relevant company, working area of the company, how they use

artificial intelligence in the company, and the market of the company is given above in Fig. 5. 2. Since artificial intelligence is a male-dominated field, no special information was given regarding the gender of the interviewees. In addition, since the number of employees in all companies is between 1-50, company scales are not included in the list. For a clearer understanding, the jobs of my interviewers and what they are doing at work are also stated.

An AI researcher from a security company describes his business as “Our company is a security company that utilizes image processing to detect fake videos generated with artificial intelligence”. One of the pressing issues in their field is that it is corresponding another artificial intelligence. The images they process for safety are outputs of different artificial intelligence.

Another security firm employee, a machine learning engineer, reports that “the firm concentrates on detecting malware files with neural networks to secure computers, which conclusively proposes next-generation security”.

Another interviewer, a sales and marketing director of an Office automation firm, affirms that “As a company, we use artificial intelligence in the textual elements. We process the data from the texts and execute comparisons. Hence, our field is text analysis for businesses”. The company offers artificial intelligence products as a helper for offices expecting to boost efficiency.

A worker from a retail solution firm who is a product manager states that “we are dealing with image processing in our company. We present statistics to partnerships by anonymizing the data of individual visitors in their stores”.

Following interviewers have several products of AI in their profession. A company’s founding partner and general manager, who work on two different domains as regulatory technologies and content management systems, explains their use of AI from several points. He states that

> "We use artificial intelligence in several areas. One of them is natural language processing. Include entity names, personal data, etc., in texts. We try to capture special patterns and then anonymize them. We also do sentiment analysis and analysis for comments in natural language processing. Another section we use is to try to make predictions by taking data from sensors. Network congestion analysis, computer vision and predict maintenance are other fields we use".

A founder and chief analyst of an advanced analytics solution firm say, "we live in a world where the correlation between different columns can be seen on structured data, and new derivative variables are produced". He asserts that they mostly use machine learning in the company, and their field is service for insurance, finance, and banking sector. After stating there are two primary purposes of their company, he explains to them as:

> "One of them is an analytical consultancy, that is, we establish categorization and prediction models based on data. The other is to develop products using algorithms. Using generalized models, we calculate the number of risks of past damages, accidents, and the different variables in them with decision trees".

He explicitly pronounces that using artificial intelligence for variables is the main point and illustrates as

> "An insurance company looks at historical data in its portfolio. How much damage women do to men, marital status, etc. He looks at them, and then he looks at them diagonally. Of course, these are easy to find when you are only male or female. Nevertheless, it is necessary to look at 60 variables simultaneously and explain the effects of the variables on each other. For subjects that are difficult to detect with the naked eye, artificial intelligence helps distinguish them."

Their second mission is to develop products where these types of information are analyzed and reported quickly.

My last interviewer is an expert on AI in the sector. He states that "Our company is a consulting company that manages advertising on social media and search engines. We do not have a direct artificial intelligence service, but all of the tools we engage are using artificial intelligence algorithms". Additionally. he is an academician in the communication field.

After expressing about the fields of study, one of the main issues to understand the Turkish ecosystem is discovering the constraints. The limitations they report and the way they evaluate limitations also acknowledge deducing for their perception of and approaches to bias.

All interviewees highlighted resources as problems or limitations. References given for resources; data, software and hardware, manpower, and processing power. It seems that the lack of data reveals as the reason for the bias, and the lack of resources can easily be the cause of technical biases, especially temporal bias. A machine learning engineer said, "There are two significant lacks for us. The first one is big data. We have limited data, and less data means more bias. Furthermore, our resources have limits. Our hardware and software resources are not adequate," and added,

> "It takes an immense number of resources to build a new model. Hardware, software resources, and people are needed. Therefore, we continue through the present literature. Sometimes we apply models directly, sometimes we run them and investigate them experimentally, and sometimes we blend various models".

Pointing out the trouble of developing a new model or algorithm, the engineer also declared how technical deficiencies prevent progress in operational processes and the association with market outcomes. The interviewer, who is the product manager, said,

> "Labor resource and technical deficiencies are a severe problem for us. Our resources are very few. Our biggest problem is processor power. We have many images available through the cameras of the stores; however, getting a cloud-based service, getting a good processor, processing power, and capacity are keeping us back. Because our data is growing exponentially".

He pointed out that while the lack of investment means not accessing sufficient resources, these deficiencies will influence the outcomes. The chief analyst also pointed to the need for resources by saying, "The biggest problem in the Turkish ecosystem is the resource. No one is aware of the investment that must be made for artificial intelligence. Human resources, operational and organizational processes are required". In conclusion, most people from the AI sector in Turkey declare the main problem as both economic and technical infrastructure.

### 5.3.1 You can trust artificial intelligence as much as the data you have

An overlook to the Turkey ecosystem confirms that one of the most prominent difficulties of the businesses is data. Within the ecosystem, data is perceived as a substantial foundation for bias. The central point is that artificial intelligence is all about processing data to create meaningful and high-performance algorithms. So, if data is not qualified enough, creating meaningful associations, and proposing stimulating outcomes becomes

more troublesome and harder. One interviewer highlighted the big data and affirmed, “Data is evaluated in five dimensions; the most valuable ones among these five dimensions are its quality and size.” However, in the lack of big data, many companies are trying to produce meaningful and high-accuracy algorithms from very limited data. Within this struggle, an extra problem arises as one interviewer pointed “models are trained with existing data, and when encountered with dissimilar data in the real world, our models may not be able to recognize the data and present results.” So, deficiency of data appears in outcomes as faulty conclusions. If one considers inside artificial intelligence reasoning, the problem arises within the structure as:

> “The main logic of artificial intelligence is to develop a model that fits your hypotheses. Then you run this model, refactoring it with real-world data and feedback. It is already impracticable to evaluate today with outdated or outdated data sets”.

Simply put, if the data is not comprehensive enough, the results are not desirable. For satisfying outcomes, the standards for data are crucial. Standards and metrics for healthy data are discussed in the following chapter.

### 5.3.2 Garbage in Garbage Out

Garbage in, garbage out is a term that refers to if data is not clean, results also cannot be clean. Artificial intelligence results base on inputs and outputs will correlate with the received data. In this sense, three points are referred to in the interviews. First, clean, and noisy data; secondly, homogeneity and representativeness of data, and societal bias are reflected in the data.

To start with, clean and noisy data concepts are essential. These concepts are also strongly related to big data and its 5V's. One interviewer explained the data-related problems as

> "The biggest problems we face in artificial intelligence are related to data. We can divide them into several headings. First, sometimes the information in the incoming data may not be apparent; we call them noisy data. It can be complicated to reach beneficial results with noisy data. Noisy data is a prevalent problem in incoming sensor data or documents containing text. The second is that, unlike noisy data, when you train the model with immaculate data, it cannot understand the different things it encounters in the real world. The difference between the real data and the training data affects the results. So, in general, data is not suitable for big data. The more you can develop the training dataset, the more successful you will be".

He pointed the problem within both clean and noisy data and the importance of big data. However, another interviewer puts more solid explanations for data criteria and why it is crucial. He acknowledged that "In other words, the biggest problem we have is that the data must comply with various criteria and standards for the algorithm to work correctly. Models look at historical data, and if there is no historical data, they cannot produce any results or benefits. At least a few years' data is a great need.

Moreover, this data must be of high quality, that is, to enter information on different variables to produce statistical results. One interviewer stated that “If your data is missing, there is nothing you can do". At first look, lack of data can be seen as more troublesome, and however, on the other hand, noisy data has more problematic results in the real-world environment. The difference lies in the fact that missing data can be realizable. Machine learning engineer said "Cleaning up data is a difficult process. Looking at our problem, we are trying to understand what is missing from our data, and we are trying to balance the data to ensure stable learning of AI". So, the lack of a dataset is not helpful yet not harmful, and it can be noticeable with the results. However, inaccurate data can result in disastrous data. One interviewer suggested that "The data is very dirty. The deficiencies in the data are something else; at least it can be decided whether to use it or not. Nevertheless, if the data is wrong, the inaccuracy of the data and its consequences will be realized much later". And these consequences are generally related to bias.

> "Incoming data biased data, what we call artificial intelligence, also works statistically. It gives results according to the weight in the data. That is why there is a definition of garbage in, garbage out in artificial intelligence. The data must be healthy and clean. Otherwise, the result is not neutral".

In other words, unhealthy data strongly means biased outcomes. The clean data concept also suggests the inclusion of differences, high representation, and homogeneity in the data. "The more different data you have, the richer your model will be. However, if your data is wrong or unbalanced, both the model and the results start to deviate. The more homogeneous the data, the more unbiased the result". So, the balance of data from several aspects is quite essential. One exemplified the importance of balance as "we were once working with a brand that mainly focused on children's clothing, but also served adults. Since most of the data came from children and the weighted data was children, the model

had a hard time detecting adults". The data was not balanced, and the representation of adults was not enough; the data concluded biased results.

AI researcher pointed the matter that "There will be less bias if your data has greater representation and overall representative distribution. Models grasp learning most easily. Therefore, it will be better if you have the most general and representative data". It is hard for the machine to understand data and put new correlations unless the data is given within the training model. Another example can be seen as "we decided not to include Covid data when Covid started. However, the problem is that after covid, the data has changed. The machine gave results with old data, but these did not work". So, it is a priority to get valid and correct data to artificial intelligence for decision making. the reason is "there is mathematics at the core of artificial intelligence is statistical data. Artificial intelligence looks at the incoming data, sees the high probability from this data, draws conclusions, and decides". The expert on the AI sector defined AI outcomes as “a data analytics provided by a very strong computer that analyzes big data” and listed questions for datasets as “what is in the data, what is the sample size, how it was selected, which methodological approaches were used in its selection?”. Briefly, what is included in the data and how it is included hold crucial points. One also emphasized the importance of classification in the data, which can cause biased results.

> "The preparation of the data is as important as the homogeneity of the data. If you change the labels, your results will also change. One of the problems with dataset bias is classifications. There are categorical features in the data, and it is very normal for bias to appear in the data when these are not cleared".

On the other hand, one stated that classifications are essential for homogeneity, yet stated that classification also has an important place for the perception of correlations.

> "What we call bias in the data sense is caused by the relationships that trigger each other and lead each other somewhere, which we call multicollinearity. For example, the relationship between weight and height for the BMI index. These two data are related and biased with each other. Serious statistical tests are required for AI results. You must make something out of the data you have. Therefore, the more limited the data you have, the more biased the result will be".

So, the data's lack explicitly ends with the bias results because data has intercorrelations, and homogeneity is not entirely something that models confront in real-world data. One stressed that

> "The homogeneity of the data, the categorization in the data is of course for some significant bias. However, I think the real problem is that algorithms are constantly running on clean data. Because real-life data is not that clean, we use categorization to reveal the heterogeneous distribution in the data. If everything were homogeneous, there would be no need for an algorithm anyway. The hard part of the job is making the heterogeneous data homogeneous within itself, which is the goal of cluster algorithms and segment algorithms. Models work correctly in homogeneous sets, but real life is messy and heterogeneous. Therefore, the data should be homogenized to be accurate".

In conclusion, although little differences can be seen in the discussion on data, the content of the data from several aspects has a correlation with biased results.

### 5.3.3 Implications for Technical Bias

As suggested in Fig. 3, one of the bias types is a technical bias that principally relies on the power of the computers or time-related effects such as temporal bias. Technical bias does not based on content, and it is more described with the technical details or timeliness. So, by modest adjustments, technical bias can be controlled smoothly. One interviewer, a machine learning engineer, affirmed that

> "We use models from the industry. There are ready-made models in the literature that proved its performance. These models are trained with valid datasets. We decide on the model by checking at the metrics and performances and, if necessary, make adjustments or adaptations to the model".

Therefore, in the case of no revision of data, temporal bias can surface in AI. It is not straight related to data itself, and it is more relevant to the fact that the changes in the society and behavior of people can conclude as negative outcomes in the case of old data sets.

Another implication on the technical bias is more explained by the lack of investment in the firms. In the absence of enough investment, companies are stuck with feasible tools which can be unequipped with the requirements of AI. So, their preferences depend on what they have more than what they need. A product manager stated that

> "When we choose a model, we first look at what we have. What power do we have? Where are these models used, what are their limits, what are the requirements for performance? Among these models, you need to find the one that suits your hardware and human resources. At the same time, the extent to which the relevant models have proven themselves is also decisive. You decide accordingly".

The decision is based on the popularity of possibilities and capacities of corporations. He also indicated that "Turkey cannot be a pioneer in model production anyway within these

circumstances. No budget to breakthrough". The infrastructure of the sector causes technical insufficiencies for the companies and affects the outcomes of models indirectly.

### 5.3.4 Responsibility of the Individual

One of the main questions is that: "Who is responsible for the bias of AI?" Without even asking the question, with the point that data is crucial for creating bias, most interviewers implied that the individual who arranges the data set or the educate model is the one. Ethical responsibility of bias associated with people. One interviewer denoted that

> "Human is the most influential factor in AI bias. The person who trained the model and the person who prepared the data. If anything is neglected in AI decisions, you need to study again the data given and provided by individuals. In other words, if it is not clean, it is due to the person who prepared the data or the programmer's irresponsible behavior during the training process".

Additionally, some referred that people who are responsible for data do not care about bias, speaking,

> "I do not think the people who prepared the dataset pay much attention to representation or homogeneity. The balance of the data determines the bias, yet I am not sure how much consideration is given or to what extent details are regarded by the person preparing the dataset. I do not think this is given much importance".

Nonetheless, there is obvious stress that in Turkey's ecosystem, attention to a dataset is not a priority or even probable. The absence of data and resources concludes with the fact that no consideration for data cleaning.

> "You do not have much luck when collecting data sets anyway. We usually must make do with the data we find. There are tremendous dilemmas for the ecosystem in Turkey. Efforts are being made to eliminate the bias in the datasets, but we do not have a study on that. We already find the dataset hardly. It is not realistic to deal with the bias of the found dataset".

It is necessary to think within the structural restrictions to understand how bias is supervised in the Turkey ecosystem. A founder declared that

> "The responsibility for bias falls on the person who created the data to some extent. Nevertheless, on the other hand, this is a very labor-intensive job. We receive unstructured data, and we try to make sense with a statistical weave. There is not much opportunity and time to clean the data".

Although cleaning data is a way to perform unbiased results, it is too straightforward that resources are moderately limited. "How you separate the data and how you extract it is

vital. The people who created the data have a responsibility here" and emphasized, "If you want a very sophisticated model, you must parse the data a lot. However, who will parse it, to looking which source, these are separate questions". Taking everything into account, individuals that are related to the AI operation process have a responsibility according to interviewers.

### 5.3.5 All is performance

To understand the bias of AI and the positions of companies towards AI, it is essential to first understand the motivations for using AI. The factors influencing the usage of artificial intelligence can obliquely designate whether measures are taken to prevent discrimination of artificial intelligence or to what extent discrimination can be allowed. In addition, although the relationship between the goals for use and discrimination is not recognized as a primary determinant, studying the relation between the reason for use, benefit and precaution structurally will provide an opportunity to contemplate the solution of discrimination in artificial intelligence. In this section, the logic of use and precautions for bias will be revealed.

The interviewers from various fields of the artificial intelligence sector answered the logic of use and metrics to determine the model uniquely. One of the most technical answers for the logic of metrics is given by the chief analyst as

> "For specific problems, hypotheses, or suggestions, there are algorithms approved for use in the literature. There are some well-known methods to follow. When deciding on a model, we begin with existing literature and approaches from literature, and at the end, relevant metrics are preferred. Minimum square of error, Gini index[xxi], standard deviation, validation matrix are some metric formats. Which metric is final or severe depends on your own business."

He proposed that firmly there cannot be a regular metric to apply. It depends on the reason for using the model. The problem that AI will solve manages the metrics of the model. He illustrated it as,

> "For instance, you are in the health sector, and you work for recognizing Covid-19 cases. The fundamental thing you must do here is to prepare your hypothesis precisely. If you have an unlimited supply of vaccines, you try to minimize false-positive patients rather than avoid over-vaccination costs. However, if you have very few vaccines and you need to apply the vaccine very accurately to the people who need the vaccine most urgently. Then you try to increase the real positive metrics".

So, metrics distinguish from each other, and they mean distinctive strategies and calculations.

The appropriate metric is selected according to the hypothesis. On the other hand, another interviewer said straight, “We define ourselves as a productivity company. That is why the main metric we use in artificial intelligence is to produce efficiency. We also pay regard to the accuracy, but our central metric is efficiency”. As a company focused on one field, it is simple to establish the metric. As well as a security firm employee said, “The central metrics we use in AI evaluation are a low margin of error and speed. We require accurate returns and demand to get them fast. The decision-making metrics of our models are based on these two concepts”. Yet, can diversity and difference in the answers be understood as just an outcome of practicing in the field of artificial intelligence? Or can different answers propose identical rationalizing? For example, one interviewer who operates with three different artificial intelligence models emphasizes that

> “We use several metrics for AI. Our principal metric in image processing is accuracy. Accuracy delivers the success rate of the business. If our results are not accurate, there is no real-world equivalent. We have a second construction where we forecast sales. Here, our principal metrics are accuracy and consistency. Consistency among data is as valuable as accuracy. Our last metric is customer satisfaction on the axis of the suggestions we offer to customers. We do not care if the suggestion is right or wrong. When customers implement our recommendations, they give us feedback, and we look at which metrics we have increased for our clients in these recommendations”.

With all the many answers, one detail is fundamental to highlight. All these companies work in different areas, and their needs are entirely different. They use artificial intelligence for various reasons; still, it is obvious for all of them, “artificial intelligence enables us to see and process data that would not normally be seen.” Their objective for the business in particular terms can be different; though, they all depend on the model for its performance. As one interviewer explained articulately,

> “Our main metric in the use of artificial intelligence is, of course, performance. After all, if you prefer artificial intelligence over anything that can be done with human intelligence, the reason is that artificial intelligence performs better. In common, you get results faster, with less margin of error, within the framework you teach artificial intelligence. If you can reach the performance you are trying for, accuracy, efficiency, they are already coming after the performance”.

Performance is essential for bias because most companies reexamine their data if the outcomes are not prosperous. If their target performance is not reached in the true-to-life world results of artificial intelligence, they review distinct components of models. At first

glance, it can be seen as ordinary; though, the actual problem is that the only technique to identify bias is the failure in the performance. One question that directs how companies recognize bias and take precautions for bias responded as

> "We try to detect bias in our tests by scanning for any problems/errors in the returns. When you export the model to the real world after training, you encounter low accuracy rates. So, something in the real world is not equal to your training data. That signifies you must check your training data. In these controls, if we notice a problem in the data, we rebuild the model".

Another representative stated that "We do not have particular rules for detecting bias. Usually, we try to fit the model to the most common and detect bias in real-life data. If we cannot detect it, seldom the customer reports and informs us" and pointed the fact that "Still, there is no specific way, method, or approach to bias". It is also explicit that there is no procedure to detect bias unless some malfunction in the decisions. Only one company employer indicated that

> "No one in the company is accountable for the bias check and taking precautions. Some do this alongside their work – machine learning engineers, for example. However, the frequency is usually once a month, at first. It is checked for a deviation in the model; it is checked for a while- one or two months-but then stops. Bias is not a problem that can be our priority".

As it can be seen in his words, even it appears like there is a detection method, it is temporary or only to check performance. Nevertheless, there is no methodology if the results are successful, even though artificial intelligence exhibits bias. The more specific question, "What is your precaution for bias? Are there any team or responsible people for detection of bias?" is answered as followings:

> "Honestly, our company does not have a process to prevent bias or take action, and we do not have the resources to spare. We do not get much feedback about bias anyway. If the model results harm the company somehow, only in these circumstances, bias becomes on the agenda. In such a case, we evaluate the reason and adjust the model or data. Orientation in the model, changing the data classifications, balancing the data, or removing the imbalance in the data may be options in such a case. On the one hand, it should not be forgotten that the importance of what artificial intelligence misses is also predominant".

> "There are areas where we try to make the data as homogeneous as possible. However, we need a team dedicated to data extraction, a team that will look at the data distribution, clean it up, and put it into the model. Finding it is one thing; maintaining it is another. Allocating resources for data cleaning and data analysis also means much money. In other words, a large team is needed to homogenize the data. Nevertheless, homogenizing the data is not among the priorities."

So, all these words, detecting bias requires resources such as workforce, algorithmic model, etc., and bias becomes an issue if there is some problem on the side of companies.

Even in these circumstances, companies do not take responsibility and leave the responsibility to the clients. If it is not a problem for the client or outcomes, detection of bias or methods to prevent bias needs so many resources, which also signifies capital.

> “When bias occurs, it also reflects on the model. We leave an automatic reporting and compare with the month the model was established and the previous month. How much change is there? We investigate it. You both detect continuity, and over time behaviors change because data is fluid. Previous models and later ones are different. We are moving forward with model comparison methods, and we are trying to give something based on both the model and its variables. But based on responsibility, we leave this situation to the customers themselves”.

Companies do not take any responsibility to prevent bias.

### 5.3.6 Inherited Bias

Talking about data invariably closes with its relationship to society. Dataset bias is not something that can be assessed outside society. The bias in the data is a representation of what society has. "The evaluation of artificial intelligence is about the data that comes to it. Therefore, it is about the experiences of the society, the individuals in the society” and stated how it happens as “People experience things, record what they have encountered, and these operate over artificial intelligence as data. The bias of artificial intelligence exhibits the bias in society". Interviewers have addressed strong cases. One example given by the product manager was that

> "We were doing natural language processing in the company while using social media data, especially Twitter. When you analyze data from one of the polarized wings – right or left – when Twitter is processing data, the way and capacity of AI to understand data from the other wing varies drastically. For example, when we examined the data on the left with the data we received from the right-wing, the model was constantly warning and expressing dissatisfaction. The models we trained with data from the left-wing understood irony better than the model we trained with the data from the right-wing. There is an incredible amount of bias in text data in particular".

Another one heavily implied the societal bias and asserted,

> "The bias in the data arises from societal biases. For example, a student at Stanford University had a graduation project. In the project, he developed a model that analyzes traffic accidents and makes loans respectively. The model studied the history of accidents and lawsuits in the data. As a decision, model granted women much more risky offers than men".

So, artificial intelligence significantly produces outcomes according to the delivered data, and the received information is already biased. One said that "If there are problems with these in the real world, there are also data. Whatever it is, there is bias in it, and it is

moving from the real world to data. How much bias can you solve through the model is also a question". Another also indicated the artificial intelligence learning depends on the past. "Because artificial intelligence learns from the past, it is biased because it feeds on social judgments and norms." Everything that society has, such as biases, norms, traditions, and behaviors, is all coded in the data acquired by artificial intelligence. As well, these data have enormous needs of quality. Concerning the societal bias, some referred that

> "There is also the discrimination side of this bias. Of course, these are sensitive issues. Here is the same problem. The results are also discriminatory, as companies do not keep a wide variety of in-depth data on their history. For example, let us take men and women. The results are as follows, and there is such a relationship between women and damage. This is explicitly and certainly discrimination. Another common cause of serious damage may be geographical features, the vehicles' quality, or the driving courses. However, there is no information about the great damage except demographic data, so we can go this far with the information we have."

Exclusion of data as well data has the societal bias settles with discriminative actions. "We prepare datasets from life, and there is bias in life. No matter how social consciousness is developed, there is bias in every society today" and he stated that "The bias and discrimination in society, we cannot overcome it. Of course, we observe bias in datasets because we cannot overcome them socially and cannot break away from these prejudices". The expert also declared that "There will always be some mistaken projections as the dataset is based on experiences and past". In the end, it is accepted by interviewers that there is a strong connection between societal bias and AI bias.

# 6. CONCLUSION

In the last years, artificial intelligence is frequently tasked with decision making roles in multiple practical domains in societal, organizational, and personal lives. The AI-based decision-making process has applied in many fields, from juridical to social media and from medical to human resources in businesses. While machine-based decisions adopted in significant societal impact spheres, level of fairness and trust have become fundamental subjects, studies show that individuals and groups representing minorities or communities about gender, race, etc., are confronted with harmful consequences, and AI produces the identical fallacious bias and discrimination. The report discovers "that big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace" (Barocas et al. 2017). As well, AI can contribute to online invisibility (Bucher 2012), inherits the existing biases and normalize social biases (Osoba and Welser IV 2017).

In this study, my hypothesis is that the bias of AI is social and associated with the socioeconomic system. I tried to show that social structure-based judgments are contained by data and transferred to artificial intelligence with data, and it formed as bias in AI. Within the framework of this context, I first associated the concept of bias with the discussion of discrimination, which is widely found in the literature. While explaining that the decisions expressed as discrimination are related to real-world outcomes and occur because of different types of bias in artificial intelligence, on the other hand, I made a new categorization of artificial intelligence bias and pointed out the relationship between data set bias and socially existing bias. I tried to show that this context must be read in the context of social shaping of technology theory, and the interaction of technology and social order. In this context, I studied the Turkish ecosystem by taking the case and interviewed 7 different interviewers.

With this research and taking the Turkey ecosystem as an instance, all the information acquired through interviews revealed several points. The first one is that people in the sector identify the bias and discrimination of AI. Although implications for what bias and discrimination are, differentiate, and in particular, these two phrases are intertwined, there is no refusal of bias. The acceptance of bias is an unquestionably observed result.

The following issue is that there is a difference in comprehending and discussing the bias of artificial intelligence. For several, both notions have sociological explanations, and some of them separate notions as a sociological topic and technical detail.

The third assumption is that bias on AI is recognized in real-world data, while the reasons for bias base on technical or data-related predicaments. Technical details are also pointed to as infrastructural necessities, which can be resolved with investments or the advancement in the sector. On the other hand, data-related problems have different viewpoints. As a technical bias, dataset problems can also be associated with infrastructure problems such as insufficient data, absence in gathering data, and not corresponding to big data standards. An extra perspective that is relevant to big data further has deductions under different descriptions. Representation in data, homogeneity, and cleanness of data are the primary references for dataset bias. People in the sector propose that noisy data is a powerful reason for bias because these data are not valid in terms of unbiased information. Poor or unequal representation and heterogeneity in the data produce bias because from the beginning, the model is trained on a biased set, and AI decisions are impressed by the variables of models within training data. Classifications in the model and data directly influence the approach of the model to real-world data. The third perspective which is vital for data bias is societal bias. Societal bias automatically provokes bias because the data appearing in the model already has prejudices, judgments, and attitudes inherited from society.

The fourth outcome is that AI bias can be diminished by individuals responsible for preparing datasets or training models. In other words, the ethical position of the person can be a determinant for bias. This aspect can be considered in two ways. First, paying attention to the data or model for bias can be dismissed by individuals. On the other hand,

the prejudices of an individual can be transferred to the model itself. Within two conditions, ethical accountability for the bias is assigned to the individuals.

The last issue is that although there is bias definitely, companies do not take responsibility to lessen bias. All interviewees pronounced that to reduce bias, having a specialized workforce is a fundamental need; however, no one is worrying about bias or employing the appropriate and requisite workforce due to the cost of workforces. Any attempt to overcome bias happens under two circumstances: low performance of AI or problems in the outcomes. So, if biased outcomes of AI affect the business's profit, then checking bias becomes a matter.

Many scholars suggest methods to diminish bias in AI. The studies focus on cleaning dataset or modifying the learning methods. Feldman et al. (2015) formed a test for disparate impact and methods by which data might be made unbiased. Situated algorithms that based on a sociotechnical systemic approach have been presented (Draude et al. 2019). Luong, Ruggieri, and Turini (2011) introduced a system for detecting discrimination by adopting classification by practicing in a historical perspective. Another learning algorithm for appropriate classification by formulating fairness advice(Zemel et al. 2013). Zafar et al. (2017) suggested that disparate mistreatment in binary classification assignment can be particularized concerning several misclassification criteria such as false-positive rates, false-negative rates, and false discovery rates. The adversarial learning method[xxii] by anonymizing data is another method for unbiased results.

Regarding all these methods to lessen bias, one of the obstacles with these problems is that they all focus on the technical part of the bias. As suggested in the research, a narrow perception strengthens bias as “technical detail”; however, a comprehensive overview exposes other factors. The fact lies in the “social construction of Artificial Intelligence”. AI’s association with society can not only be degraded the conception of “social product”. With the fact that as technological development, AI is a technical design produced by humans, on the other hand, AI’s relation with the socio-economic conditions is a

continuous active for AI outputs. Two aspects define what AI does and how AI carries bias:

The first one is the societal bias inherited by AI. Under societal bias, again, two points are fundamental subject. The existed bias of society is transferred to the AI through data. It can be thought of in terms of people and community. To put other words, existed knowledge and attitude of society from juridical to human resources all operates into the data. For both people who generate data and run the model and give AI data without even knowing it, the individuals' choices cause bias. Bias, prejudices, judgments, discriminatory behaviors, all easily can reflect in AI and appears in the decision-making process as discriminative actions.

Besides inherited bias, the second major point for bias is that the economics of the sector determines the approach to the bias. It is strongly affirmed through interviews that knowing bias and having the possibility to diminish bias does not indicate attempts to prevent bias. The economics, in the age of capitalism, the market conditions become the basic determinant for assessing bias. The profit-based system weakens ethical attitudes.

So, my point of view is that artificial intelligence bias must be discussed in terms of socio-economic systems and capitalism. The market competition policies and profits created by capitalism also pursue its superstructure or in other words, life generally.

> "economic structure of society, is the real basis on which the juridical and political superstructure is raised and to which definite social forms of thought correspond; that the mode of production determines the character of the social, political, and intellectual life generally" (Marx 1859, 2).

If the relationship between economic base and consciousness becomes clear, it will be easier to understand how societal bias strongly connected to economic structure.

> "The production of ideas of conceptions of consciousness is at first directly interwoven with the material activity and the material intercourse of men, the language of real life. Conceiving, thinking, the material intercourse of men appears at this stage as the direct efflux of their material behavior. The same applies to mental production as expressed in the language of politics, laws, morality, religions, metaphysics, etc. of a people. Men are the producers of their conceptions, ideas, etc. – real active men, as they are conditioned by the development of their productive forces and the forms of intercourse corresponding to these, up to its furthest forms. Consciousness can never be anything else than conscious existence, and the existence of men is their actual life process."(Marx and Engels 1970, 36)

Consciousness is the subjective expression of objectively existing relations. It emerges as a consciousness of participation in these relations.

> "Ideas and thoughts of people, then, are ideas and thoughts about themselves and of people in general…for it [is] the consciousness not merely of a single individual but of the individual in his interconnection with the whole of society"(Marx and Engels 1970, 83)

In this perspective, without socio-economic transformation, it is impossible to develop unbiased outcomes from AI.

To put it in other words, the solution for bias of AI cannot be found in technical developments. Technology is shaped by social settings. The bias of AI and its discriminatory behaviors cannot be thought outside the social shaping. The socio-economic conditions within all technological development become the primary determinant of how technology is used. As David Noble (2017) stated ", technology is not the problem, nor is it the solution. The problem is political, moral, and cultural, as is the solution: a successful challenge to a system of domination which masquerades as progress" (351). This thesis has tried to show that bias is not just a subject that belongs to numerical sciences and is stuck in technical discussions, but that social sciences are an area that should be examined economically, sociologically, and culturally.

As conclusion remarks, with all these implications and questions, how AI can be used for society and people is one of the main problems and indicates hope for the future. Overall, AI is produced by people, yet its operation is based on the reality of capitalism, with the aim of profit rather than providing a field for the common interests of people.

## REFERENCES


‘Adversarial Machine Learning Definition | DeepAI’. n.d. Accessed 28 June 2021. https://deepai.org/machine-learning-glossary-and-terms/adversarial-machine-learning.

Aysolmaz, Banu, Deniz Iren, and Nancy Dau. 2020. ‘Preventing Algorithmic Bias in the Development of Algorithmic Decision-Making Systems: A Delphi Study’. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

Baeza-Yates, Ricardo. 2018. ‘Bias on the Web’. *Communications of the ACM* 61 (6): 54–61.

Baltzan, Paige. 2013. ‘Business Driven Technology 6th Edition’.

Barocas, Solon, Elizabeth Bradley, Vasant Honavar, and Foster Provost. 2017. ‘Big Data, Data Science, and Civil Rights’. *ArXiv Preprint ArXiv:1706.03102*.

Barocas, Solon, and Andrew D Selbst. 2016. ‘Big Data’s Disparate Impact’. *Calif. L. Rev.* 104: 671.

Barr, Avron, and Edward A. Feigenbaum. 1981. ‘The Handbook of Artificial Intelligence. William Kaufmann’. *Inc., Los Altos, CA*, 1.

Baştan, Serhat. 2017. *İletişim Teorisi: Teknolojik ve Kültürel Yaklaşımlar Üzerinden Bir Eylem ve Yapı Çözümlemesi*. Ankara: Orion Kitabevi.

Baxter, Leslie A, and Earl R. Babbie. 2003. *The Basics of Communication Research*. Cengage Learning.

Bell, Daniel. 1972. ‘The Cultural Contradictions of Capitalism’. *Journal of Aesthetic Education* 6 (1-2): 11–38.

Bini, Stefano A. 2018. ‘Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?’ *The Journal of Arthroplasty* 33 (8): 2358–61.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. ‘Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings’. *ArXiv Preprint ArXiv:1607.06520*.

Bozdag, Engin. 2013. ‘Bias in Algorithmic Filtering and Personalization’. *Ethics and*

*Information Technology* 15 (3): 209–27. https://doi.org/10.1007/s10676-013-9321-6.

Bucher, Taina. 2012. 'Want to Be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook'. *New Media & Society* 14 (7): 1164–80.

Buolamwini, Joy, and Timnit Gebru. 2018. 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification'. In *Conference on Fairness, Accountability and Transparency*, 77–91. Proceedings of Machine Learning Research.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases'. *Science* 356 (6334): 183–86.

Cam, Arif, Michael Chui, and Bryce Hall. 2019. 'Global AI Survey: AI Proves Its Worth, but Few Scale Impact'. https://www.mckinsey.com/featured-insights/artificial-intelligence/global-ai-survey-ai-proves-its-worth-but-few-scale-impact.

Chaturvedi, Devendra K. 2008. 'Soft Computing'. *Studies in Computational Intelligence* 103.

Ciampaglia, Giovanni Luca, Azadeh Nematzadeh, Filippo Menczer, and Alessandro Flammini. 2018. 'How Algorithmic Popularity Bias Hinders or Promotes Quality'. *Scientific Reports* 8 (1): 1–10. https://doi.org/10.1038/s41598-018-34203-2.

Cobb, Helen G, and Peter Bock. 1994. 'Using a Genetic Algorithm to Search for the Representational Bias of a Collective Reinforcement Learner'. In *International Conference on Parallel Problem Solving from Nature*, 576–87. Springer.

Copeland B J. 2020. 'Artificial Intelligence | Definition, Examples, and Applications | Britannica'. 2020. https://www.britannica.com/technology/artificial-intelligence.

Crawford, Kate. 2021. *The Atlas of AI*. Yale University Press.

Dalenberg, David Jacobus. 2018. 'Preventing Discrimination in the Automated Targeting of Job Advertisements'. *Computer Law & Security Review* 34 (3): 615–27.

Danks, David, and Alex John London. 2017. 'Algorithmic Bias in Autonomous Systems.' In *IJCAI*, 17:4691–97.

Dastin, Jeffrey. 2018. 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women | Reuters'. Reuters. 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Datenethikkommission, Die, Der Begriff, Das Eckpunktepapier, and Ki- Strategie. 2018. 'Empfehlungen Der Datenethikkommission Für Die Strategie Künstliche Intelligenz Der Bundesregierung', No. September: 1–5.

Daumeyer, Natalie M, Ivuoma N Onyeador, Xanni Brown, and Jennifer A Richeson. 2019. 'Consequences of Attributing Discrimination to Implicit vs. Explicit Bias'. *Journal of Experimental Social Psychology* 84: 103812.

Demchenko, Y, C de Laat, and P Membrey. 2014. 'Defining Architecture Components of the Big Data Ecosystem'. In *2014 International Conference on Collaboration Technologies and Systems (CTS)*, 104–12. https://doi.org/10.1109/CTS.2014.6867550.

Dietterich, Thomas G, and Eun Bae Kong. 1995. 'Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms'. Citeseer.

Dobbe, Roel, Sarah Dean, Thomas Gilbert, and Nitin Kohli. 2018. 'A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics'. *ArXiv Preprint ArXiv:1807.00553*.

Draude, Claude, Goda Klumbyte, Phillip Lücking, and Pat Treusch. 2019. 'Situated Algorithms: A Sociotechnical Systemic Approach to Bias'. *Online Information Review*.

Emspak, Jesse. 2016. 'How a Machine Learns Prejudice'. *Scientific American, December* 29: 2016.

Ennals, Richard. 2016. 'Artificial Stupidity'. *AI & SOCIETY* 31 (3): 431–32.

'Etkinlikler - Türkiye Yapay Zekâ Inisiyatifi'. n.d. Accessed 17 October 2021. https://turkiye.ai/etkinlikler/.

Feenberg, Andrew. 2002. *Transforming Technology: A Critical Theory Revisited*. Oxford University Press.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. 'Certifying and Removing Disparate Impact'. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–68.

Foucault, Michel, and François Ewald. 2003. *' Society Must Be Defended': Lectures at the Collège de France, 1975-1976*. Vol. 1. Macmillan.

Frana, Philip, and Michael J. Klein, eds. 2021. *Encyclopedia of Artificial Intelligence:*

*The Past, Present, and Future of AI*. ABC-CLIO.

Friedman, Batya, Peter H Kahn, and Alan Borning. 2008. 'Value Sensitive Design and Information Systems'. *The Handbook of Information and Computer Ethics*, 69–101.

Fürnkranz, Johannes. 2010. 'Decision Tree BT - Encyclopedia of Machine Learning'. In , edited by Claude Sammut and Geoffrey I Webb, 263–67. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_204.

Garbade J Michael. 2018. 'Understanding K-Means Clustering in Machine Learning | by Dr. Michael J. Garbade | Towards Data Science'. 2018. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre. 2017. 'Correlation and Variable Importance in Random Forests'. *Statistics and Computing* 27 (3): 659–78.

'Hakkımızda - Türkiye Yapay Zekâ Inisiyatifi'. n.d. Accessed 28 June 2021. https://turkiye.ai/hakkimizda/.

Haraway, Donna. 1988. 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective'. *Feminist Studies* 14 (3): 575–99.

Harnad, Stevan. 2008. 'The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence (PUBLISHED VERSION BOWDLERIZED)', 1–28. http://eprints.soton.ac.uk/262954/1/turing.html.

Heath, Nick. 2018. 'What Is Deep Learning? Everything You Need to Know | ZDNet'. 2018. https://www.zdnet.com/article/what-is-deep-learning-everything-you-need-to-know/.

Hempel, J. 2017. 'Melinda Gates and Fei-Fei Li Want to Liberate AI from "Guys With Hoodies"'. WIRED. 2017. https://www.wired.com/2017/05/melinda-gates-and-fei-fei-li-want-to-liberate-ai-from-guys-with-hoodies/.

Herbert, J Rubin, and Irene Rubin. 1995. 'Qualitative Interviewing: The Art of Hearing Data'. Thousand Oaks, CA: Sage Publications, Inc.

HLEG ECAI. 2019. 'A Definition of Artificial Intelligence: Main Capabilities and Scientific Disciplines'. Brussels. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

Howard, Ayanna, and Jason Borenstein. 2018. 'The Ugly Truth about Ourselves and Our

Robot Creations: The Problem of Bias and Social Inequity'. *Science and Engineering Ethics* 24 (5): 1521–36.

Hunt, Elle. 2016. 'Tay, Microsoft's AI Chatbot, Gets a Crash Course in Racism from Twitter | Artificial Intelligence (AI) | The Guardian'. Guardian. 2016. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.

IBM Cloud Education. 2020. 'What Is Supervised Learning? | IBM'. Ibm. 2020. https://www.ibm.com/cloud/learn/supervised-learning.

Joshi, Ameet. 2020. *Machine Learning and Artificial Intelligence*. Springer.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Macmillan.

Kaski, Samuel. 2010. 'Self-Organizing Maps BT - Encyclopedia of Machine Learning'. In , edited by Claude Sammut and Geoffrey I Webb, 886–88. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_746.

Katyal, Sonia K. 2019. 'Private Accountability in the Age of Artificial Intelligence'. *UCLA L. Rev.* 66: 54.

Kirchner, Lauren, Surya Mattu, Jeff Larson, and Julia Angwin. 2016. 'Machine Bias — ProPublica'. Propublica. 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Kvale, Steinar. 1996. *Interviews : An Introduction to Qualitative Research Interviewing / Steinar Kvale*. Thousand Oaks, California: Sage Publications.

Larson, Erik J. 2021. *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. Belknap Press: An Imprint of Harvard University Press.

Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard university press.

Leavy, Susan. 2018. 'Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning'. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, 14–16.

Lee, Nicol Turner. 2018. 'Detecting Racial Bias in Algorithms and Machine Learning'. *Journal of Information, Communication and Ethics in Society*.

Lee, Nicol Turner, Paul Resnick, and Genie Barton. 2019. 'Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms'. *Brookings Institute: Washington, DC, USA*.

Lerman, Jonas. 2013. ‘Big Data and Its Exclusions’. *Stan. L. Rev. Online* 66: 55.

Lu, Huimin, ed. 2021. *Artificial Intelligence and Robotics*. Springer.

Luong, Binh Thanh, Salvatore Ruggieri, and Franco Turini. 2011. ‘K-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention’. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 502–10.

Marwala, Tshilidzi. 2021. *Rational Machines and Artificial Intelligence*. Academic Press.

Marx, Karl. 1859. ‘Preface to a Contribution to the Critique of Political Economy’. *The Marx-Engels Reader* 2: 3–6.

Marx, Karl, and Friedrich Engels. 1970. *The German Ideology*. Vol. 1. International Publishers Co.

Mayson, Sandra G. 2018. ‘Bias in, Bias Out’. *YAle LJ* 128: 2218.

McCarthy, John. 2007. ‘What Is Artificial Intelligence’.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. ‘A Survey on Bias and Fairness in Machine Learning’. *ArXiv Preprint ArXiv:1908.09635*.

Mermin, N David. 2007. *Quantum Computer Science: An Introduction*. Cambridge University Press.

Mester, Tomi. 2017. ‘Statistical Bias Types Explained (with Examples)’. Data36. 2017. https://data36.com/statistical-bias-types-explained/.

Murphy, Laura. 2016. ‘Airbnb’s Work to Fight Discrimination and Build Inclusion A Report Submitted to Airbnb’. *Airbnb.Com*. http://blog.airbnb.com/wp-content/uploads/2016/09/REPORT_Airbnbs-Work-to-Fight-Discrimination-and-Build-Inclusion.pdf.

Naylor, C David. 2018. ‘On the Prospects for a (Deep) Learning Health Care System’. *Jama* 320 (11): 1099–1100.

Nilsson, Nils J. 1998. *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. nyu Press.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. ‘Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries’. *Frontiers in Big Data* 2 (July): 13. https://doi.org/10.3389/fdata.2019.00013.

Oppermann, Artem. 2020. *What Is Deep Learning and How Does It Work? - Towards Data Science*. *Medium*. https://towardsdatascience.com/what-is-deep-learning-and-how-does-it-work-2ce44bb692ac.

Osoba, Osonde A, and William Welser IV. 2017. *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Rand Corporation.

Press, Gil. 2017. 'Alan Turing Predicts Machine Learning And The Impact Of Artificial Intelligence On Jobs'. 2021 Forbes Media LLC. 2017. https://www.forbes.com/sites/gilpress/2017/02/19/alan-turing-predicts-machine-learning-and-the-impact-of-artificial-intelligence-on-jobs/?sh=7426f9261c2b.

Roetzel, Wilfried, Dezhen Chen, and Xing Luo. 2020. 'Genetic Algorithm - an Overview | ScienceDirect Topics'. Design and Operation of Heat Exchangers and Their Networks. 2020. https://www.sciencedirect.com/topics/engineering/genetic-algorithm.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

Sammut, Claude, and Geoffrey I Webb, eds. 2010a. 'Bayesian Network BT - Encyclopedia of Machine Learning'. In , 81. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_65.

———. , eds. 2010b. 'Case-Based Learning BT - Encyclopedia of Machine Learning'. In , 147. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_96.

———. , eds. 2010c. 'Relational Value Iteration BT - Encyclopedia of Machine Learning'. In , 862. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_722.

Simon, Herbert A. 2019. *The Sciences of the Artificial*. MIT Press.

'Startuplar - Türkiye Yapay Zekâ Inisiyatifi'. n.d. Accessed 17 October 2021. https://turkiye.ai/girisimler/.

Strauß, Stefan. 2018. 'From Big Data to Deep Learning: A Leap towards Strong AI or "Intelligentia Obscura"?' *Big Data and Cognitive Computing* 2 (3): 16.

Suresh, Harini, and John V. Guttag. 2019. 'A Framework for Understanding Unintended Consequences of Machine Learning'. *ArXiv Preprint ArXiv:1901.10002*, 208–15.

Tennery, Amy, and Gina Cherelus. 2016. 'Microsoft's AI Twitter Bot Goes Dark after Racist, Sexist Tweets'. Reuters. 2016. https://www.reuters.com/article/us-

microsoft-twitter-bot-idUSKCN0WQ2LA: Accessed Dec 2018.

Thornhill, John. 2021. 'Is AI Finally Closing in on Human Intelligence_ _ Financial Times'. *Financial Times Magazine*. https://www.ft.com/content/512cef1d-233b-4dd8-96a4-0af07bb9ff60.

Tyagi, Neelam. 2020. 'Understanding the Gini Index and Information Gain in Decision Trees | by Neelam Tyagi | Analytics Steps | Medium'. 2020. https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8.

Uther, William. 2010. 'Temporal Difference Learning BT - Encyclopedia of Machine Learning'. In , edited by Claude Sammut and Geoffrey I Webb, 956–62. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_817.

Wilner, Alex S. 2018. 'Cybersecurity and Its Discontents: Artificial Intelligence, the Internet of Things, and Digital Misinformation'. *International Journal* 73 (2): 308–16.

Wischmeyer, Thomas, and Timo Rademacher. 2020. *Regulating Artificial Intelligence*. Springer.

Woosley, Lynn W., and Max B. Sherman. 2019. 'Big Data, Machine Learning, and Bias'. *The RMA Journal* July-Aug. https://rmajournal.org/rmajournal/july_august_2019/MobilePagedArticle.action?articleId=1504204#articleId1504204.

'Yapay Zeka Kullanımı Raporu'. n.d. Accessed 28 June 2021. https://www.keyofchange.com/tr/2224/Yapay zeka kullanımı raporu/.

Yudkowsky, Eliezer. 2008. 'Artificial Intelligence as a Positive and Negative Factor in Global Risk'. *Global Catastrophic Risks* 1 (303): 184.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. 'Fairness beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment'. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–80.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. 'Learning Fair Representations'. In *International Conference on Machine Learning*, 325–33. Proceedings of Machine Learning Research.

# APPENDIXES

## APPENDIX A: Various Versions of Interview Questions

1) How do you use artificial intelligence in your professional life/company?
2) What are the limitations that you encounter in the operation of AI?
3) What are the problems that you encounter in the operation of AI?
4) How do you assess AI operation such as effective, profitable, productive etc.?
5) Can you give a brief information about AI operation process?
6) What are the factors that affect AI decision-making process?
7) Which metrics are deterministic in AI's decisions?
8) Which metrics do you use when you choose modelling for your AI training?
9) Do you think AI has bias? If you say yes, why AI has bias?
10) Do you think that there is relationship between AI's assessment and societal biases?
11) If you believe that AI's assessment is related to societal biases/judgements, can you describe the relationship in a detailed way?
12) Do you think AI recognizes particularities?
13) Do you have multiple sources of data?
14) Is your data set comprehensive?
15) Do you believe your dataset has enough representation of people?
16) Do you think the person that is responsible from structuring data has any effect on AI's bias?
17) To what extent your institution's policy affect AI's operation?
18) If you realize that there is bias in your AI operation, what cautions do you take? How you manage the process?
19) What proportion of your resources is appropriate for an organization to devote to assessing potential bias?
20) Who leads in your organization's effort to identify bias in its AI systems?

## CURRICULUM VITAE

**Kişisel Bilgiler**

Adı Soyadı :Şeyda Tuğgen Gümüşay

**Eğitim Durumu**

Lisans Öğrenimi :Kadir Has Üniversitesi / Yeni Medya

Yüksek Lisans Öğrenimi :Kadir Has Üniversitesi/ İletişim Bilimleri ABD- Yeni Medya

Bildiği Yabancı Diller :İngilizce

**İş Deneyimi**

Çalıştığı Kurumlar ve Tarihleri:İstanbul Ticaret Üniversitesi (2019-devam ediyor)

**İletişim**

:

## NOTES

[i] The discussion on behavior of AI is given in chapter 3.2.1, 3.2.4, and the Fig.3.1.
[ii] Defining mimic or reproduce differs in detail, however for the thesis it is not vital as much.
[iii] Case-based learning refers to a family of techniques for classification and regression, which produce a class label/predication based on the similarity of the query to its nearest neighbor(s) in the training set (Sammut and Webb 2010a).
[iv] A Bayesian network is a form of directed graphical model for representing multivariate probability distributions (Sammut and Webb 2010b).
[v] A decision tree is a tree-structured classification model, which is easy to understand, even by nonexpert users, and can be efficiently induced from data (Fürnkranz 2010).
[vi] Linear regression is used to identify the relationship between a dependent variable and one or more independent variables and is typically leveraged to make predictions about future outcomes (IBM Cloud Education 2020).
[vii] K-means clustering identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible (Garbade J Michael 2018).
[viii] A computational data analysis method which produces nonlinear mappings of data to lower dimensions (Kaski 2010).
[ix] Neural networks open up a feature-rich framework with practically unlimited scope to improve the performance for the given training data by increasing the complexity of the network (Joshi 2020, 50)
[x] Genetic algorithms are randomized search algorithms that have been developed to imitate the mechanics of natural selection and natural genetics. Genetic algorithms operate on string structures, like biological structures, which are evolving in time according to the rule of survival of the fittest by using a randomized yet structured information exchange (Roetzel, Chen, and Luo 2020).
[xi] Deep learning algorithms attempt to draw similar conclusions as humans would by continually analyzing data with a given logical structure. To achieve this, deep learning uses a multi-layered structure of algorithms called neural networks (Oppermann 2020).
[xii] A generalization of the dynamic programming technique for solving Markov decision processes (MDPs) that exploits the symbolic structure in the solution of relational and first-order logical MDPs through a lifted version of dynamic programming (Sammut and Webb 2010c).
[xiii] A method for computing the long term utility of a pattern of behavior from a series of intermediate rewards (Uther 2010).
[xiv] A form of TD Learning
[xv] An artificial intelligence field involves mining human text and speech to generate or respond to human inquiries in a readable or ordinary way. NLP has required advancements in statistics, machine learning, linguistics, and semantics to decode natural human language's uncertainties and opacities (Frana and Klein 2021, 245).
[xvi] It is from Arxiv, page number is not stated in website.
[xvii] A behavioral bias type: exposed as syntactic, semantic, and structural differentiations in the content generated by users.
[xviii] Behavioral bias that is manifested as differences in the qualities of networks collected from user associations, interactions or actions.
[xix] Volume, velocity, variety, value and veracity (Demchenko, Laat, and Membrey 2014).
[xx] It is from website, so page number is not given.
[xxi] Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class then it can be called pure (Tyagi 2020).
[xxii] Adversarial Machine Learning is a collection of techniques to train neural networks on how to spot intentionally misleading data or behaviors. This differs from the standard classification problem in machine learning, since the goal is not just to spot "bad" inputs, but preemptively locate vulnerabilities and craft more flexible learning algorithms ('Adversarial Machine Learning Definition | DeepAI' n.d.)