HELLENIC REPUBLIC
National and Kapodistrian University of Athens

National and Kapodistrian University of Athens

Department of History and Philosophy of Science

&

Department of Informatics and Telecommunications

Interdepartmental Graduate Program:

Science, Technology, Society—Science and Technology Studies

MSc Thesis

"Big Tech and AI Ethics: Inputs from STS"

Konstantinos Konstantis

Registration Number: 17/220

Thesis Advisory Committee:

Gotsis Georgios, Professor (advisor)

Tympas Aristotle, Professor (member)

Drakopoulos Stavros, Professor (member)

Athens 2021

Abstract

In this thesis, I claim that artificial intelligence (AI) ethics cannot be adequate and complete if it does not include Science and Technology Studies (STS) approaches; I suggest STS inputs that should be included in AI ethics. AI is everywhere: AI systems take decisions for us constantly. Users of AI have started to be worried about the consequences that AI has in their lives. Thus, tech companies, citizens and governments have started to talk about AI ethics. Privacy, transparency, biases, accountability and inclusiveness are only some of many ethical problems that arise as a result of AI. In this thesis, I will first provide a historical contextualization for AI. Second, I will offer an overview of the ethical problems which come together with AI, and which users and designers of AI should be aware of. Third, I will give an STS perspective on AI, in terms of how it is designed by big tech and how this perspective impacts AI ethics; in my view, talking about AI ethics without taking into account STS approaches is not enough. Fourth, I will refer to the secondary literature on AI ethics and the way that the media presents big tech and its relation with AI ethics. Fifth, I will analyze the policy of Microsoft in relation to AI ethics. As I will explain in Chapter 3, among big tech companies, Microsoft has one of the most complete and adequate policies in the field of AI ethics. It is my intention to point out that even Microsoft's policy in AI ethics, however, requires an STS perspective.

Keywords: AI (artificial intelligence) ethics and STS, big tech, Microsoft's policy and AI ethics, transparency/black box and AI, bias and AI

Περίληψη

Σε αυτή τη διπλωματική, ισχυρίζομαι ότι η ηθική της τεχνητής νοημοσύνης δεν μπορεί να είναι επαρκής και πλήρης εάν δεν περιλαμβάνει προσεγγίσεις από το επιστημονικό πεδίο STS. Για το λόγο αυτό, προτείνω προσεγγίσεις από το επιστημονικό πεδίο STS, οι οποίες θα πρέπει να περιλαμβάνονται στην ηθική της τεχνητής νοημοσύνης. Η τεχνητή νοημοσύνη είναι παντού: Τα συστήματα τεχνητής νοημοσύνης λαμβάνουν αποφάσεις για εμάς συνεχώς. Οι χρήστες της τεχνητής νοημοσύνης έχουν αρχίσει να ανησυχούν για τις συνέπειες που έχει η τεχνητή νοημοσύνη στη ζωή τους. Συνεπώς, οι τεχνολογικές εταιρείες, οι πολίτες και οι κυβερνήσεις έχουν αρχίσει να μιλούν για την ηθική της τεχνητής νοημοσύνης. Η ιδιωτικότητα, η διαφάνεια, οι προκαταλήψεις και η λογοδοσία είναι μόνο μερικά από τα πολλά ηθικά προβλήματα που προκύπτουν ως αποτέλεσμα της τεχνητής νοημοσύνης. Σε αυτή τη διπλωματική, αρχικά θα παραθέσω μια ιστορική πλαισίωση της τεχνητής νοημοσύνης. Δεύτερον, θα προσφέρω μια επισκόπηση των ηθικών προβλημάτων που συνδέονται με την τεχνητή νοημοσύνη και τα οποία πρέπει να γνωρίζουν οι χρήστες και οι σχεδιαστές της. Τρίτον, θα παρουσιάσω μία οπτική από το επιστημονικό πεδίο STS για την τεχνητή νοημοσύνη, όσον αφορά το σχεδιασμό της από τις μεγάλες τεχνολογικές εταιρείες και τον τρόπο με τον οποίο αυτή η οπτική επιδρά στην ηθική της τεχνητής νοημοσύνης. Κατά την άποψή μου, μία συζήτηση για την ηθική της τεχνητής νοημοσύνης που δεν περιλαμβάνει προσεγγίσεις από το επιστημονικό πεδίο STS δεν μπορεί να είναι επαρκής. Τέταρτον, θα αναφερθώ στη δευτερογενή βιβλιογραφία για την ηθική της τεχνητής νοημοσύνης και στον τρόπο με τον οποίο τα ΜΜΕ παρουσιάζουν τις μεγάλες τεχνολογικές εταιρείες και τη σχέση τους με την ηθική της τεχνητής νοημοσύνης. Πέμπτον, θα αναλύσω την πολιτική της Microsoft σε σχέση με την ηθική της τεχνητής νοημοσύνης. Όπως θα εξηγήσω στο Κεφάλαιο 3, μεταξύ των μεγάλων τεχνολογικών εταιρειών, η Microsoft έχει μια από τις πιο ολοκληρωμένες και επαρκείς πολιτικές στον τομέα της ηθικής της τεχνητής νοημοσύνης. Πρόθεσή μου είναι να επισημάνω ότι ακόμη και η πολιτική της Microsoft ωστόσο, στον τομέα της ηθικής της τεχνητής νοημοσύνης, απαιτεί μια οπτική από το επιστημονικό πεδίο STS.

Λέξεις κλειδιά: Ηθική της Τεχνητής Νοημοσύνης και STS, μεγάλες τεχνολογικές εταιρείες, πολιτική της Microsoft και ηθική της τεχνητής νοημοσύνης, διαφάνεια/μαύρο κουτί και τεχνητή νοημοσύνη, προκατάληψη και τεχνητή νοημοσύνη

Table of contents

Table of figures

Acknowledgments

First of all, I want to thank my supervisor, Professor Georgios Gotsis, for his guidance throughout my research for this master's thesis. Professor Gotsis helped me both in terms of pointing me toward the right literature to read and suggesting the right corrections to perfect the content and structure of thesis. Moreover, I would like to thank Professor Stavros Drakopoulos for his participation in the three-member committee examining my dissertation.

I would also like to thank Manolis Simos, a postdoctoral fellow in the department of history and philosophy of science, who played a crucial role in helping me understand ethics and the way that it combines with technology. Manolis also gave me specific and valuable advice on the completion of my thesis.

I would also like to thank Professor Aristotle Tympas very much for the confidence that he has shown in me. After I graduated from the school of electrical and computer engineering, Professor Tympas gave me the opportunity, through this interdisciplinary master's and PhD program, to discover the world of STS and develop a passion for it. His previous experience as an engineer means that he can always understand the way I think and guide me with the right advice. He is the most significant factor in my achieving an STS perspective, but always with the view of an engineer.

Finally, I firmly believe that without the help in all areas from my family and the people close to me, I would not have been able to complete this thesis.

1. Introduction

In this thesis, I argue that artificial intelligence (AI) ethics, in its current form, is not adequate and requires a Science and Technology Studies (STS) perspective to be truly complete and achieves the goal of benefiting the whole of society. AI is all around us: AI systems are being used in almost every area of our lives, such as finance, social media, health, manufacturing, etc. In many cases, AI systems have entered our lives so completely that we do not even notice that there is an AI system doing the work. Alongside this domination of AI in our everyday lives has come the domination of big tech in the field of AI. Although, in theory, anyone could develop an AI system with the right knowledge and a computer, in fact AI systems are developed by big tech companies.

But before discussing big tech and its policy in AI ethics, in this thesis I will offer a historical contextualization of AI and a review of AI ethics. To do this, I will study, among other articles, one by Colin Garvey (2018), "Broken Promises and Empty Threats: The Evolution of AI in the USA, 1956-1996", for its historical contextualization, and one by Vincent Müller (2020), "Ethics of Artificial Intelligence and Robotics" (in the *Stanford Encyclopedia of Philosophy*), for its review of AI ethics. AI cannot be discussed as a neutral technology developed autonomously and separately from the social and economic environment; as Garvey analyzes in his article, the development of AI should be examined in the context of the spectrum of its interactions with scientists, governments and society in general. As he explains, the growth of AI is not linear, in contrast to the history often presented for it. In fact, AI has been through "booms" and "winters". Booms come with funding for AI, based on promises or threats, and winters with economic depression; the "false alarm" of promises and threats will be analyzed in Chapter 2.1. After the historical contextualization of AI, I will refer to AI ethics. More and more people are starting to be affected by AI, and more and more ethical issues have arisen as a result. Privacy and surveillance, manipulation of behavior and the biases that come from AI are only a few of these ethical issues. The article "Ethics of Artificial Intelligence and Robotics" also discusses the issues that arise if AI systems are considered subjects. The "morality" and "conscience" of robots are only two of many such issues. Other scholars, however, such as De Cremer and Kasparov (2021), disagree with the approach of seeing AI systems as independent from their designers.

I believe that it is important to see AI ethics from a different perspective and discuss this following the overview of AI ethics. Using a book by Cathy O'Neil (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, and an article by Jenna Burrell (2016), "How the machine 'thinks': Understanding opacity in machine learning algorithms", I will try to offer STS inputs for AI ethics. One of the most significant inputs is the way that Burrell defines *opacity*. This definition, which is different from most of the definitions of opacity used by big tech companies,

allows for an explanation of the creation of a "black box" and its role. I refer to "black box" as an imaginary box that has technology in it and where people cannot see its operation, its function and the way that this technology comes to an output. An STS approach includes the way that a technology – in this case, AI – is designed and the role of this design. As part of this framework, I will investigate the economic, social and political role that the design – and more specifically, the "black box" that results from this design – can play. I will use O'Neil's work to refer to many cases of AI systems where tech companies suggest something went wrong, leading to ethical issues; in fact, O'Neil claims that nothing went wrong and AI systems did their job as they really are, i.e., "weapons of math destruction," a phrase that will be analyzed in Chapter 2.3.

As we will see, the role of big tech in the field of AI, and therefore in the field of AI ethics, is significant. After the discussion of the STS approach, I will present articles from the secondary literature that criticize (from their own perspectives) the way that big tech is concerned with AI and the way that it develops AI ethics. Big tech companies develop AI ethical guidelines in their quest for ethical AI, but many citizens and scholars—such as Sam Gilbert (2020) in his article "Big Tech and Data Ethics"—criticize them for "ethical washing". There are also many articles such as "Community-in-the-loop: towards pluralistic value creation in AI, or—why AI needs business ethics", by Häußermann and Lütge (2021), supporting the position that current policies by big tech companies in the field of AI ethics are not enough to benefit society; these suggest that different principles and ethical values should be embedded in AI ethics, such as *order ethics*, as Häußermann and Lütge suggest.

Next, I will refer to the media due to its significant and dual role in the field of AI. On the one hand, the media informs citizens about the development of AI by big tech companies and ethical issues that have arisen; on the other hand, big tech may face pressure by the media and be forced to develop better and more adequate policies in AI ethics.

As I will show in this thesis, Microsoft seems to have an adequate and complete AI ethics policy. However, even Microsoft's policy seems to be insufficient due to the lack of an STS perspective. I will analyze Microsoft's policy in AI ethics through the primary literature of Microsoft's papers, and how its tries to avoid biases and discrimination, be transparent, develop fair algorithms and use unbiased data. Then, in conclusion, I will refer to specific examples of Microsoft's policy and suggest that what is missing is an STS approach. Without this approach, and without STS inputs, Microsoft and all other tech companies cannot have a truly adequate policy in AI ethics, one that would benefit the whole of society.

2. Contextualizing AI: History, ethics and STS

In the beginning of this chapter, I will offer a historical contextualization of AI. To do that, I focus on Garvey's (2018) paper "Broken Promises and Empty Threats: The Evolution of AI in the USA, 1956-1996". According to him, AI can be separated into three periods – or AI "booms". These booms come from promises that cannot be kept and threats that do not really exist. Between these booms, there are AI "winters", when funds for AI are limited.

After the historical contextualization of AI, I will offer an overview of AI ethics. Using the article "Ethics of Artificial Intelligence and Robotics" that appears in the *Stanford Encyclopedia of Philosophy*, I will mention the ethical issues that arise alongside AI, such as privacy, employment and singularity, and will focus on *opacity*, as the most significant theme to investigate from an STS perspective.

Finally, investigating AI and AI ethics from an STS perspective, I will study the article "How the machine 'thinks': Understanding opacity in machine learning algorithms" by Burrell (2016), which gives an STS perspective on the operation of machine learning algorithms. I will also provide STS inputs into the field of AI, based on O'Neil's book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.

2.1 AI: Historical contextualization

First, I will approach AI from an STS perspective, using Garvey's article "Broken Promises and Empty Threats: The Evolution of AI in the USA, 1956-1996" because this article analyzes one of the best arguments for the point I want to highlight. As Garvey (2018) explains, AI is not a neutral technology developed without being affected by the social, economic and political environment in which it belongs. This is an STS perspective requiring a historical contextualization of AI; it is necessary for the whole concept of STS inputs in AI ethics due to the *neutrality*, *autonomy* and *opacity* of AI. Garvey (2021, p. 1) refers to AI as a "suite of techniques intended to make machines capable of performing tasks considered 'intelligent' when performed by people" in his article "Unsavory medicine for technological civilization: Introducing 'Artificial Intelligence & its Discontents'". In his article "Broken Promises and Empty Threats: The Evolution of AI in the USA, 1956-1996", Garvey (2018) analyzes AI with a completely different approach from the one taken by most narratives about it; these, as Garvey claims, have been written by scientists who are close to the technology. As a result, the history focuses on the benefits of AI rather than the risks and the mistakes that have been made. He gives special weight to the impact that statements about the future of AI have had. According to him, due to these statements, which have the form of promises and threats, and more

specifically the form of broken promises and empty threats, there have occurred three AI "booms" and, between them, two "winters". As Garvey shows, there were multiple promises made for AI in its booms, shaping the reality through funds and expectations, but mostly these have not been kept.

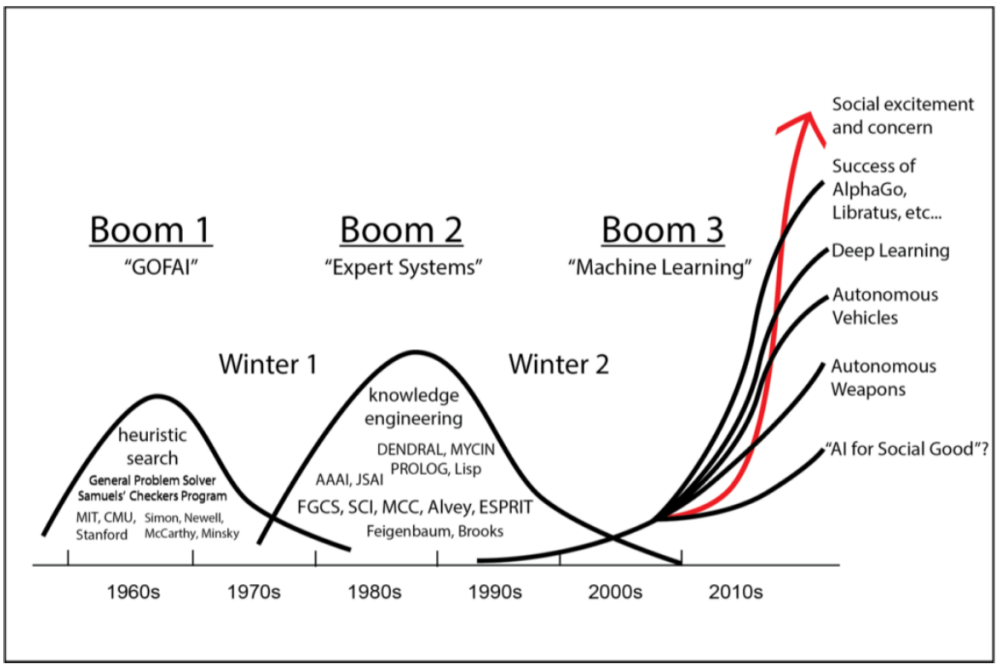

Figure 1. The three booms of AI, according to Garvey (original diagram inspired by Yutaka Matsuo) (source: "Broken Promises and Empty Threats: The Evolution of AI in the USA, 1956-1996", Garvey, 2018)

According to Garvey (2018), the first AI boom was in the late 1950s and early 1960s. After World War II, there was true competition between the USSR and the USA. Due to this competition, a lot of money was provided to the military for the development of technology. In this era, AI pioneers John McCarthy, Marvin Minsky, Herbert Simon and Alan Newell dominated in this field. Garvey focuses on Simon and Newell, because they played a leading role in making promises shaping the future of AI. Two years after a landmark conference at Dartmouth College, the first to use the term "artificial intelligence", and one year after the launch of Sputnik by the USSR, the USA founded the Advanced Research Projects Agency (ARPA) to provide funds for R&D. In this period, Simon and Newell presented their expectations of AI to a meeting with the agency; ARPA immediately offered to fund research into AI. Yet the predictions that these two scientists made were over-optimistic. They predicted that a digital computer would have the ability to be a chess champion before the 1970s. They also insisted that "there are now in the world machines that think, that learn, and that create" (Garvey, 2018, p. 4). There is no necessity for someone to live almost 60 years later to understand that these statements were over-confident in the late 1950s. As Garvey says, Richard Bellman, a popular scientist, very soon expressed his concerns about these statements, saying that "they were not scientific writing" (Garvey, 2018, p. 5). Simon and Newell answered that as scientists they had to tell people the truth, which is why they made such ambitious statements. As a result, they gained extensive funding for their research from ARPA. There were also statements from the pioneers of AI

about the risks of "cybernation" – a word that comes from "cybernetics" and "automation" – which comes with AI. Despite the fact that in the early 1960s there were too many concerns, both from scientists and governments, about "cybernation" and the risks of it for the labor force, by the late 1960s everyone had realized that these statements would not become true. What Simon and Newell had promised could not be accomplished. Therefore, ARPA decided to cut funding for most AI projects; in the late 1960s, as Garvey says, the first AI winter began (Garvey, 2018, pp. 3–7).

This "winter" ended in the late 1970s when the second AI boom started. Edward Feigenbaum, a student of Simon, played a leading role in the second boom. Feigenbaum understood that promises could not bring funding to AI research any longer and focused on other ways to achieve that. He had a plan with both a technical and a social part. For the former, Feigenbaum focused on "expert systems" (Garvey, 2018, p. 7). Instead of the general intelligence which most AI scientists proposed up till then, he believed that systems should be experts in specific areas. For the latter, he focused on the idea of the threats of AI instead of the promises made in the first boom. To accomplish his goal, Feigenbaum found a perfect opportunity with the rise of Japan in the technological field. His purpose was to convince the American government, ARPA and industry in general to fund AI R&D, due to the threats coming from Japan. In the early 1980s, Japan started a huge program called "Fifth Generation Computer Systems" (FGCS), with the intention of rebuilding the Japanese economy after World War II. The purpose of FGCS was to produce computers and AI systems. From the late 1970s, the program was underway, and Feigenbaum was one of the scientists from outside Japan that took part in it. Yet when Feigenbaum went back to USA, he presented FGCS as a threat to America. He claimed that Japan would come to dominate both in the computer industry and in the information society that would prevail in the coming years. As a result, Feigenbaum presented FGCS as a threat to the USA, suggesting that the country should fund a project focused on AI in response to this (Garvey, 2018, pp. 7–10).

However, as Garvey (2018, pp. 10–11) highlights, FGCS was not a threat to the USA: "The project did not include plans for the production of any commercial products, much less for domination of the global economy." Feigenbaum knew this, but intentionally described the project in his own way to benefit from it. And the benefits came: Microelectronics and Computer Technology Corporation and the Strategic Computing Initiative were two of the first programs that came in answer to the "threat" of Japan. These programs were huge and provided millions in AI R&D. But this was only the beginning. They were the reason to start many more programs, such as European Strategic Programme on Research in Information Technology, focusing on AI and funding its research. And so began the second AI boom. But it came with some criticisms of AI. Joseph Weizenbaum, an AI scientist, in 1983 "equated the promise of Boom 1 with the threat of Boom 2" (Garvey, 2018, p. 12). As a result, threats became empty threats as a few years they were not fulfilled. In the late 1980s and

the early 1990s the second AI boom collapsed and the second winter of AI began, as "Japan did not dominate the globe with AI. Instead, its economy collapsed in the early 1990s. The Japanese FGCS was consequently described as a failure by the US computing community, and many rewrote history to claim they had always seen the FGCS as an empty threat" (Garvey, 2018, p. 13).

In late 1990s, when IBM's Deep Blue beat the chess grandmaster Gary Kasparov, the third "boom" begins. Despite the fact that one year earlier, in 1996, Kasparov had been the winner in the battle, almost everyone remembers his loss. Alongside this, the promise that Simon and Newell saw in the late 1950s was finally coming true. Thus, the third boom began. We are currently living through the third AI boom, with machine learning dominating, together with its promises and threats, which this time come from China. USA is already in an AI race with China. As Garvey highlights, the only thing remaining to be seen in the coming years is if all these promises fail again and if all these threats will go unrealized (Garvey, 2018).

Studying Garvey's paper "Broken Promises and Empty Threats: The Evolution of AI in the USA, 1956-1996" is vital to understanding that the political, social and economic environment plays a crucial role in the development of AI, cloaked by opacity is used differently in different circumstances.

2.2 AI ethics: An overview

The previous chapter has provided a historical contextualization of AI, based on Garvey's article "Broken Promises and Empty Threats: The Evolution of AI in the USA, 1956-1996". In this chapter, I will offer an overview of AI ethics. It will be based on Müller's "Ethics of Artificial Intelligence and Robotics" (*Stanford Encyclopedia of Philosophy*, 2020). I focus on Müller's article as it constitutes one of the most comprehensive and up-to-date accounts of the matter.

I offer this overview because I believe that before analyzing STS perspectives in AI ethics, it is vital to analyze AI ethics in its current form. This will allow us to see if AI ethics in its current form has limits and if so, what it requires to be more adequate. In other words, to criticize AI ethics, you first have to understand deeply the whole idea of it.

According to Müller (2020), AI ethics should be separated into two categories. The first category regards AI systems as objects and the second as subjects. Privacy, manipulation, opacity, bias, human–robot interaction, employment and the effects of autonomy are some of the issues that belong to the first category, while machine ethics, artificial moral agency and singularity are some of the issues that belong to the second. First, I will offer an overview of all AI ethical issues that Müller refers to, and then I will analyze the issue of opacity, because I believe it has a crucial role in the AI

ethics battle, and at the same time I think that an STS perspective on opacity may help us to understand AI and AI ethics better.

Figure 2. AI: To be or not to be? (Source: https://static.techspot.com/images2/news/bigimage/2019/04/2019-04-08-image-28.jpg)

According to Müller, concerns always arise alongside new technologies. In the case of AI, the ethics is multifaceted. Müller claims that AI ethics has become the new "green" energy and often such ethics is used by companies not for its actual usefulness, but for "ethics washing".

As Müller highlights, it is hard to establish a globalized AI ethics policy. Policy can take many forms and, in many times, may come into conflict with policy on other technologies. Laws, guidelines and frameworks for ethical AI are some attempts to control AI and subsume it under ethical values.[1] Concerning laws, as Rességuier and Rodrigues (2020) claim, ethical guidelines have limited effectiveness when they have a legal form. In their view, ethical codes and guidelines should not have the role of complementing laws. They claim that complying with the laws does not mean in all cases that a tech company is acting ethically. Ethics should have a flexible form and should always be adapted to the needs of society, while laws do not have that form (Rességuier & Rodrigues, 2020).

The first ethical issues that Müller analyzes are privacy and surveillance. It is known that AI uses data. Nowadays, all of our data has a digital form; it is very difficult to control who collects our data and for what purposes. Sensors and many other technologies are being used to turn aspects of our

[1] An example of ethical guidelines and framework is the Ethics Guidelines for Trustworthy AI from the High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission. According to AI HLEG, trustworthy AI has three dimensions. First, it has to be lawful. Second, it has to be ethical. Third, it has to be robust. The goal of this framework, according to AI HLEG, is "to maximize the benefits adopted by all people or organizations that somehow are connected with AI". For example, designers, users, developers, institutions, companies and researchers in AI could find these guidelines and framework useful. The four principles that always should be respected by everyone mentioned above are the principles of respect for human autonomy, prevention of harm, fairness and explicability. Finally, there are seven requirements working as guidelines for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing and accountability (High-Level Expert Group on Artificial Intelligence, 2019) (for more information: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai).

lives that have not digital form into digital data, so they can be used to train and test AI systems. Many people are worried about their private data and concerned about how such data might be used by big tech companies without the owner's permission. Companies exploit the "smart idea", which has started to spread everywhere, from *smart homes* and *smart cities* to *smart governance*, to manipulate and mislead people for their own advantage (Müller, 2020, pp. 9-10).

The second issue has to do with the manipulation of behavior. AI can be used in many ways for several reasons. For example, chatbots are a form of AI system. This combination of algorithms and data could be a way to manipulate people with political propaganda. Or, as Coeckelbergh (2020, p. 32) notes, AI could provide information about us in hidden ways that are not suspicious on first sight. A doll, for example, that uses AI to interact with a child could provide the private data of the parents to the company that had made it, and at the same time through talking with the child, manipulate it and shape its personality. Müller claims that it is hard to stop these many forms of manipulation that may occur through AI. As he highlights, the USA does not have a legal system adequate to prevent this. On the other hand, EU has the GDPR (General Data Protection Regulation), which is one attempt to deal with these problems.

Third is bias in decision systems. As Coeckelbergh (2020, p. 128) claims, "bias can arise in the selection of the training data set; in the training data set itself, which may be unrepresentative or incomplete; in the algorithm; in the data set the algorithm is given once it is trained; in decisions based on spurious correlations; in the group that creates the algorithm; and in wider society". Müller refers to the fact that AI systems often for our lives take decisions and often we do not even know it. AI systems may decide if someone meets the standards to be hired or to be offered a loan. AI systems are behind many decisions that may change our lives, so it is understood how important it is for them not to introduce bias.

Fourth, Müller refers to the fact that humans tend to develop emotions toward animals and even toward things. With robots that can interact with people and often look like them, it is unavoidable that people may develop feelings about them that will affect their behavior. An example is care robots, which are being used to help elderly people. The question is how the emotions that people have for machines and the emotions that people attribute to machines may alter their behavior.

Automation and employment comprise the fifth issue that Müller refers to. Often technology comes with automation. Automation means that fewer people will be needed to accomplish the same task. AI is no exception. Together with automation comes talk about the job losses and unemployment. Will AI lead to mass unemployment or will it lead to increased overall wealth? According to Müller, AI has some features that make people think that bringing justice to employment through AI is

difficult. First, accountability is hard to find in AI systems. Second, big tech – i.e., monopolies – dominate the AI market. Third, AI systems promote "intangible assets" which are difficult to control.

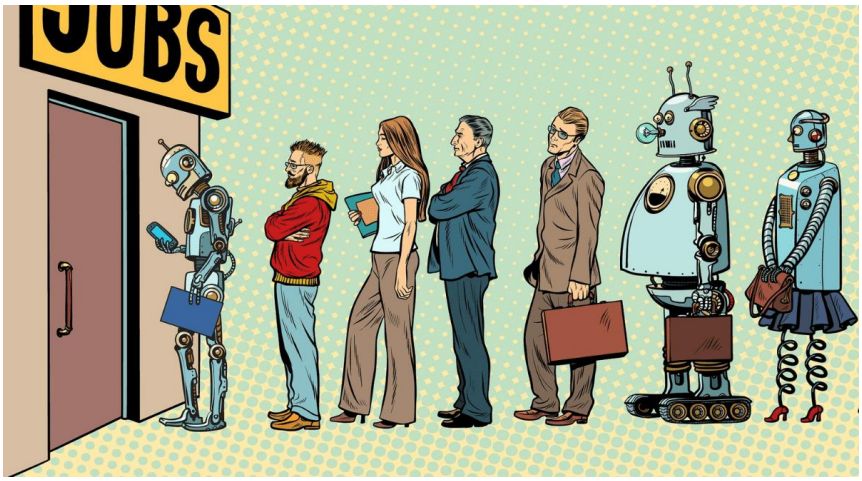

Figure 3. Living with the fear that AI will dominate in labor market (Source: https://media.hashcashconsultants.com/should-we-worry-about-robots-taking-our-jobs-in-the-future/)

Autonomous systems, e.g., autonomous cars and autonomous weapons, represent the final issue in the category "AI systems as objects" analyzed by Müller. Autonomous cars promise to solve the problem of the many deaths and injuries that arise through non-autonomous vehicles. Ethical concerns about autonomous cars arise, too. The issue of accountability cannot be answered: who is responsible for the decisions of an autonomous car and who is accountable of a lethal car accident? The "trolley problems" that capture the public attention cannot be used to solve the problems that arrive with autonomous vehicles (Müller, 2020, p. 26). The trolley problems have many forms. Müller refers to one example among many. A train is going straight ahead to five people and will kill them, unless it changes track, in which case it will kill another person. These dilemmas, according to Müller, cannot offer much to the discussion about ethical AI because are too abstract, and nobody may have to deal with them in real life. These problems cannot help us to design more ethical autonomous vehicles. In terms of autonomous weapons, the main issue is that they make killing easier. It seems that no human is responsible for a killing, because it is done by a drone or another AI system. Leaving aside the lost accountability, someone who controls such a drone does not experience killing or living under conditions of war, which may make killing and war more likely to happen.

Müller also analyzes the issues that arise when AI machines are considered as subjects. First, he analyzes machine ethics. According to him, machine ethics takes it for granted that AI robots could be moral and accountable for their own actions. According to Coeckelbergh (2020, pp. 47–54), some believe that robots could not have a conscience as people do, because they cannot think or have emotions, as people do. On the other hand, some people believe that robots could be fully moral and behave like people. They claim that robots could be more logical and take better decisions from humans because they can control their emotions. Finally, Coeckelbergh claims that there is a middle

position. Here, we could give to robots some kind of morality, but not the full morality that we accord to people. Robots should be accountable for some decisions they make, but for some others they should not.

Concerning the morality of robots, Müller (2020, p. 31) refers to them as "artificial moral agents". He suggests that if people consider them and treat them as "artificial moral agents", then such robots should have rights and responsibilities. The responsibility of robots has a complex meaning. Considering the middle position that Coeckelbergh refers to, who will be responsible for a decision if the robot is not responsible for it, bearing in mind the designers, developers, engineers and users that are involved? Except from their responsibilities, some believe that robots should also have rights, meaning that we should treat them as human beings, and not, for example, turn them off.

Finally, Müller refers to the question of singularity. According to him, if people can create machines that could have intelligence as humans do, then these machines will have the ability to create machines that surpass human abilities. However, Müller has second thoughts about the feasibility of the singularity.

As mentioned before, I will focus now on the issue that I believe is the most significant when we talk about AI and AI ethics. This is the *opacity* of AI systems. According to Müller (2020, pp. 13–16), opacity belongs to the first category of issues, the one that categorizes AI systems as objects, and aims to investigate the way that they are designed and created. Opacity may be seen from different perspectives, be caused by different factors and have multiple results. Opacity favors the absence of accountability: AI systems take decisions for our lives and at the same time we do not know who is taking these decisions. If something goes wrong or if the system is biased, the people that are affected by the system do not know who is accountable. Did the bias come from the design, from the data or from something else? Nobody knows. "Nobody" is not an exaggeration. Müller suggests that in many cases AI systems are opaque about their decisions even to their creators. As I will show in the next chapter, Burrell (2016) agrees and claims that this opacity comes from a combination of algorithms and mathematics at such a level that it is impossible for humans to understand. Müller says that opacity could be dealt with many ways; political will could transform AI into a transparent technology. This idea focuses on the fact that there are not adequate guidelines and laws to make companies and designers develop ethical and transparent AI. These ways include ethical guidelines for designing AI systems, such as the Ethics Guidelines for Trustworthy AI from the High-Level Expert Group on Artificial Intelligence mentioned above (footnote 1).

Müller also refers to other examples of criticism that have been made of the opacity of AI. One of these is the criticism that becomes from O'Neil in her book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (O'Neil, 2016). According to her, if we want to

deal with opacity in AI systems, we should investigate the way that these systems are designed. Opacity, according to O'Neil, comes from the nature of these models, as we will see in the next chapter, and could lead, intentionally or unintentionally, to biases against people.

This overview of AI ethics is needed to make comparisons with the STS approaches that will be analyzed in the next chapter. This overview is also required to see if Microsoft's policy, which will be analyzed in Chapter 3.3, is in line with AI ethics in its current form.

2.3 STS approaches to AI

I believe that the best way to understand how AI works from an STS perspective is the article "How the machine 'thinks': Understanding opacity in machine learning algorithms", by Burrell (2016). The analysis of this article is required to criticize AI ethics from an STS perspective and will be very useful at the conclusion of my thesis, where I criticize Microsoft's policy in AI ethics from such a perspective.

In her article, Burrell (2016) analyzes the problem of opacity in AI, focusing on algorithms of classification and ranking. As Burrell explains, algorithms use data as input and determine an output; most of time even the AI system's designers do not know how the algorithms reached that decision. This is the opacity of algorithms. As Burrell mentions, algorithms have already taken on broader meaning in our everyday lives. The media uses the concept of the "algorithm" to describe almost everything that takes decision, with the involvement of people or not. Yet companies have procedures where sometimes even the people involved with them do not know how decisions have been made. In this "war for transparency", which have already begun from so many perspectives, Burrell tries to give answers both from a technical and STS perspective. Burrell claims that there are three reasons causing the opacity of AI. First, opacity comes from the intention of companies to protect their patents in the field of AI. Second, opacity comes from the fact that designing AI needs special skills; not all people can "read" programming languages. Third, opacity comes from inconsistency between the way that human mind solves algorithms and mathematical problems and the way that machines do this.

Analyzing the first reason that leads to opacity, Burrell (2016, pp. 3–4) explains that this type has many perspectives. Companies and organizations may want to keep their secrets not only for the competence between them, but also for other strategic reasons. Having opacity in algorithms gives the opportunity to these corporations to handle them as they wish, without external interference. At the same time, Burrell refers to the fact that opacity could cover the intention of companies to manipulate or discriminate against people for their own benefit. The open source movement is

opposed to this idea and proposes a way to make algorithms accessible to all, without companies losing their competitiveness. Burrell believes that for the first type of opacity, the computer coding should be available for scrutiny. If this happens, corporations will not have the opportunity for violation, because everything would be accessible to everyone.

As the second cause of opacity, Burrell (2016, p. 4) refers to the fact that AI is difficult for people without special skills to design, program and read. Most people do not know how to write code in any programming language. Code that is good both for people and for machines is one that has been written in such a way that people and machines can read it, but this is not easy, because programming is difficult to explain at all stages. Burrell believes that a way to confront this cause of opacity is the existence of diversity in STEM (Science, Technology, Engineering, Mathematics) field and an attempt that must be made by those pursuing other professions, such as journalists, first to understand the algorithms and then to explain them to the public.

Finally, Burrell focuses and analyzes more on the third cause of opacity. This is the opacity that comes from the fact that algorithms can push the limits of human capabilities. This is a form of opacity that even the designers of AI have to deal with. Burrell claims that making an algorithm comprehensible for most programmers, let alone for the public, means that the algorithm may not be useful. In many cases, machine learning algorithms have to be complex to perform at their best. Due to the enormous amounts of data that are used to train and test them, it is unavoidable that algorithms will be complex. A code may be comprehensible and even large amounts of data may be manageable, but the combination of them will almost certainly create opacity. Machine learning algorithms operate on two parallel stages. In the first stage, the "classifier" takes an input and gives an output, while in the other, "'learners' must first train on test data" (Burrell, 2016, p. 5). After this training, the classifier will use its results to classify new entry data. To understand the third form of opacity, Burrell analyzes two examples of machine learning algorithms. First, she analyzes a neural network used for image recognition, and then analyzes spam filtering and shows how such filtering may cause classificatory discrimination.

Burrell (2016, pp. 5–7) focuses on a classic neural network algorithm for image recognition, which is an algorithm that can recognize handwritten digits from 0 to 9. These digits can be written in a box with 64 pixels. The neural network has an *input layer*, a *hidden layer* and an *output layer*.

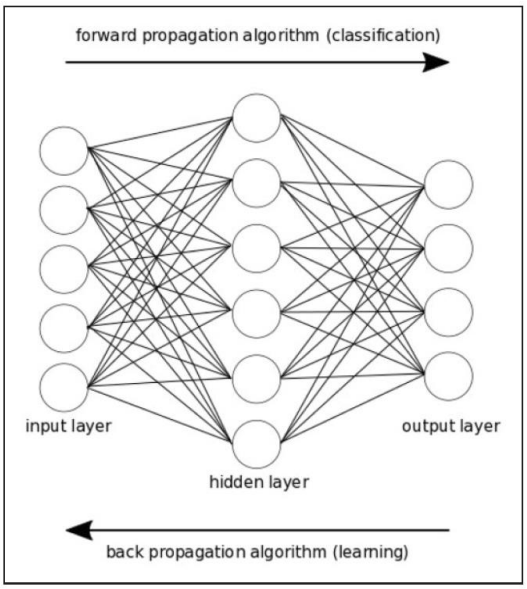

Figure 4. The layers of a neural network as described by Burrell (Source: "How the machine 'thinks': Understanding opacity in machine learning algorithms", Burrell, 2016, p. 6)

The connection between the input, hidden and output nodes forms what the algorithm will learn. The amount of black and white that each pixel has will shape the importance of each pixel for each digit. Burrell analyzes the way that hidden layer nodes work. She provides a picture (Figure 5) to understand the way that hidden layer nodes break down the problem. Each of these 25 boxes has some darker and some brighter areas. This is what each of these boxes understand as part of a digit. Each of these boxes could be one node of hidden layer and without recognizing handwritten digits as people do, with curves, lines etc., has a specific job to do: to recognize quantities of black or white in specific areas for each digit.

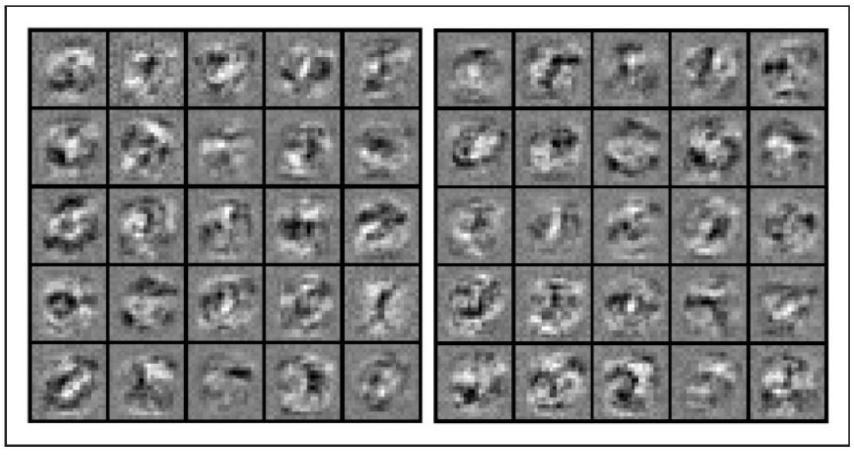

Figure 5. The hidden layer. On the left running for the first time and on the right running for the second time. The results are different "because of the random initialization step that defines the set of weights initially to very small random numbers". (Source: "How the machine 'thinks': Understanding opacity in machine learning algorithms", Burrell, 2016, p. 7)

Burrell gives us this example in her attempt to make us understand the way that machines "think". Yet, as she claims, this is not enough to understand the way that this "thinking" could lead to discrimination; she therefore moves on to the next example of machine learning algorithms.

In her second example, Burrell (2016, pp. 7–9) focuses on the way that algorithms filter spam e-mails. Spam filtering is a classification problem, giving an algorithm an e-mail (input) to decide if it is spam or not (output). Burrell investigates if there is a tendency toward labeling e-mail as spam, according to region. For example, if someone is located in areas where there is a lot of spam activity, such as West Africa or Eastern Europe, the result that his or her e-mails will be labeled as spam. Burrell analyzes Support Vector Machines, which are similar to neural networks and use a "linear kernel". In these cases, the algorithms analyze only the words of the e-mails and not the meaning of them. Every word gains particular weight according to whether this word is connected with spam or non-spam e-mails. Burrell (2016, p. 8) investigates "the Nigerian 419 scam", and wonders if the word "Nigeria" and other placenames could lead to "false positive" spam e-mails, that is, e-mails that are labeled as spam, but in fact are not. Words are graded between -1 and 1 according to their relevance to spam e-mails. If a word has been graded -1, according to the algorithm, it has nothing to do with spam. On the other hand, if a word has been graded 1, it is certain to be connected with spam e-mails. Burrell highlights that the word "Nigeria" has been graded -0.001861. So, is almost neutral in its association with spam e-mails. Meanwhile, other words like "our", "click", "visit" and "want" have a high score and are associated with spam e-mails. Burrell mentions two examples of e-mails, showing that spam filtering algorithms evaluate e-mails according to the weights that these words have, and take decisions about the validity of the e-mails, in contrast with people, who check if an e-mail is spam according to the whole meaning of the e-mail.

Scientists can often understand the reason why algorithms have evaluated some words as being strongly connected with spam e-mails, but many others they cannot. There are controversies about the fact that words such as "visit" and "want" are highlighted as dangerous. Interpretability is a concept that must not being forgotten. AI systems may not just take decisions, but could also explain how and why they reached those outputs. In fact, Google's Gmail gives the user this opportunity by providing a reason, if user wants it, for classifying an e-mail as spam. However, spam filtering is not as simple as recognizing digits, and the interpretability of it is complex. As Burrell mentions, the volume of data is constantly growing. As this happens, AI algorithms gather more data for their inputs, making the evaluation and improvement of them a more complex process, and converting the interpretability of the algorithm's decision into a more difficult task. Burrell proposes three ways to face these problems. First, machine learning algorithms may not be used in particular fields. Second, there are some methods, such as "feature extraction", to make the algorithm use only the data that is truly useful; this avoids feeding it with unrelated data. Third, Burrell (2016, p. 9) claims that "some

solutions perhaps wisely abandon answering the 'why' question and devise metrics that can, in other ways, evaluate discrimination. For example, in 'Fairness Through Awareness' a discriminatory effect in classification algorithms can be detected without extracting the 'how' and 'why' of particular classification decisions."

Burrell achieves to give an STS perspective in the opacity of AI through technical terms. Analyzing the examples of image recognition through neural networks and spam filtering through Support Vector Machines comes to the conclusion that opacity does not always come about as the intention of AI designers, but often from the representations and processes that AI systems follow to classify and take decisions. Thus, there are cases in which even AI engineers cannot understand 'why' or 'how' the algorithm made a decision. This establishes a difference between Burrell's analysis and legal scholars, for example, who demand a better legal framework to inspect the way engineers design AI. As Burrell (2016, p. 10) highlights, "alleviating problems of black boxed classification will not be accomplished by a single tool or process, but some combination of regulations or audits (of the code itself and, more importantly, of the algorithms functioning), the use of alternatives that are more transparent (i.e., open source), education of the general public as well as the sensitization of those bestowed with the power to write such consequential code".

To explore the role of opacity and therefore the role of black box in AI, I will turn to the book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, by O'Neil (2016). This book gives me the opportunity to study opacity and the black box of AI from an STS perspective, and the way that these terms are connected with biases and other ethical issues of AI.

O'Neil (2016) explains the concept of "Weapons of Math Destruction" (WMD). Investigating AI and the role of opacity, bias and black box, I believe it is vital to understand the idea of WMD, because AI includes algorithms and mathematics. O'Neil refers to many examples to explain how the combination of algorithms and mathematics is not neutral and could produce biases, especially under the cloak of a black box, which is used to represent science and technology as neutral.

IMPACT is a system which evaluates how good the teachers at a school are. It has been used in many schools to decide which of the teachers should be fired (Turque, 2011). There are cases where teachers evaluated from the principal as top in their work and who get excellent reviews from their students' parents are evaluated with a low IMPACT score. IMPACT does not take feedback into account, and does not provide feedback. As a result, it cannot improve, nor justify its decision why a teacher should be fired. But how does this opacity help teachers understand what they are doing wrong and how they could improve? Using the results of students' tests to measure teacher effectiveness, as IMPACT does, is not fair. There are so many factors that could play a role in shaping the test results of students.

For example, if a student at that time does not have a good relationship with his or her parents, he or she may fail, but this has nothing to do with the teacher's skills. IMPACT also creates a negative loop for the teachers in a non-popular school, because it automatically assumes that a teacher in a school like this is not good enough. WMDs "define their own reality and use it to justify their results" (O'Neil, 2016, p. 19). For O'Neil, this is the true purpose of a black box, which comes from the nature of WMD (O'Neil, 2016, pp. 19–21). I believe that with this example, O'Neil claims that digitalizing some of the aspects of our "analog life" will certainly lead to bias and discrimination, and will certainly lead to negative results for aspects of our lives that cannot be represented with numbers.

In the same category with the IMPACT are recidivism models. These are models that are used to decide which prisoners are likely to repeat an illegal action or not, so which should be released from prison. An example is COMPAS. O'Neil argues that these algorithms may seem neutral on race, but this is not true. They may not ask prisoners about their race, but the questions that are being asked are based on history and not the case of the prisoner itself. This will definitely lead to biases (Starr, 2014). Prisoners, for example, are being asked if they have been stopped by police in the past, and the relation of their friends with illegal actions, etc. As O'Neil claims, this is not fair. An African-American would be more likely to be stopped for a police check than a white man. As a result, the algorithm reproduces injustice arising from society. This is a typical example of WMD (O'Neil, 2016). Research by Angwin et al. (2016) came to the conclusion that the ability of such systems to predict crime is really low. In 2013 and 2014, only 20% of the people predicted to carry out a crime actually did so. The concept of the questions for evaluation is also problematic. Someone may abuse a child and get a low score because he or she may have a job; on the other hand, a drunk man may get a high score because he is homeless. African-Americans get higher scores than white people, which, on the algorithm, means that they have more chance of committing a crime again. Finally, in Northpointe, an assessment tool for the prisoners, 23.5% of white defendants labeled with higher risk did not re-offend; the percentage of African-American defendants in the same category was 44.9%. At the same time, 47.7% of white defendants who were labeled with lower risk did re-offend, while only 28.0% of African-American defendants who labeled with lower risk did re-offend (Angwin et al., 2016). According also to O'Neil (2016, pp. 34–38), recidivism models decide for a person's future not based on his or her actions but based on his or her past. But as we have seen, discrimination and bias may dominate in someone's past, and this is not his or her fault or something that he or she should be punished for. As O'Neil highlights (2016, p. 38), recidivism models do what IMPACT does and what WMDs do: they lead a person "into a pernicious WMD feedback loop" (O'Neil, 2016, p. 38).

PredPol is another system that is presented as neutral. PredPol predicts the place where a crime is more possible to be committed. It does not predict who is going to commit the crime, but the place

of the action. This might indicate that this system cannot be biased against people, but this is not true. If PredPol gives a signal of a crime in an area, regardless of the magnitude of the crime, this area will be a target in the algorithm, which will always send police to that area (Berg, 2014). The reference to the magnitude of the crime was not random. O'Neil (2016, pp. 97–104) points out that algorithms such as PredPol include not only homicides, assaults and other violent crimes, but also "nuisance" crimes such as consuming small quantities of drugs. Including crimes such as these in a system such as PredPol will lead the police to increase patrols in areas where these crimes occur, but not to encounter the most violent crimes. Also, it will lead to producing new data that will create a loop against these areas, which will justify more policing. In this way PredPol shapes reality and reproduces a loop again and again. When an algorithm is based on geographical statistics and not on human statistics, it does not mean that it is neutral. PredPol is nothing more than the use of technology to justify police patrols in underprivileged areas where more African-American people live: "The result is that we criminalize poverty, believing all the while that our tools are not only scientific but fair" (O'Neil, 2016, p. 104).

As O'Neil suggests, letting algorithms rule our lives will definitely not lead us to a fair and ethical future. We cannot know which of our actions from the present will in some way affect our future. Living in a world which everything is evaluated may lead us toward a state of competition which may make us use improper means to complete our goals. We will always be in categories – in many cases without even knowing it – about our habits, our preferences, our income etc., and we will not even know why are we in these categories and what this categorization means for our lives. A rejection for a loan or from a university will always follow someone for the rest of his or her life, and it affects how other people treat that person. Being poor or belonging to a minority is getting more and more expensive and dangerous in a world of WMDs (O'Neil, 2016).

So, it is vital for the whole society that scientists, and especially the engineers and designers of AI, are aware that it is in the nature of AI that it could create biases without the intention of its designers. Engineering students should always have in mind that technology is not neutral; trying to avoid biases is not easy as simply avoiding race or sex. School curricula should always include lessons about biases and the black box of AI. In fact, we should perceive AI as WMD under the cloak of black box, and not as something neutral that always takes the right decisions.

Figure 6. Bias in AI (Source: https://mostly.ai/2020/05/04/why-bias-in-ai-is-a-problem/)

Analyzing the article by Burrell (2016), "How the machine 'thinks': Understanding opacity in machine learning algorithms", and the book by O'Neil (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, was necessary to understand the meaning of the STS approach in AI and AI ethics. The criticism that will be made of Microsoft's AI ethics, but also of the AI ethics of big tech as a whole, will be based on this approach.

3. AI and business: The case of big tech

In the first part of this chapter, I will use the secondary literature and articles such as "Community-in-the-loop: towards pluralistic value creation in AI, or—why AI needs business ethics", by Häußermann and Lütge (2021), to carry out a review of big tech and AI ethics. Then, by studying articles from media outlets such as *New York Times*, I will show the way that the media presents big tech and its relation with AI. The media has a dual role: informing citizens and pushing tech companies for improvements. Finally, I will analyze the policy of Microsoft in AI ethics. To do so, I will use papers by Microsoft setting out its principles, guidelines, rules, etc.

3.1 Secondary literature review

In this section, I provide an overview of the secondary literature regarding the interrelations between big tech, AI and AI ethics. Specifically, I use a series of recent articles that substantiate from an AI ethics perspective the problematic way in which big tech deals with the ethical issues mentioned earlier. These criticisms, when they are not from an STS perspective, are useful for contextualizing the whole idea of big tech and AI ethics, but in fact in many cases they seem to be based on common

ideas with STS approaches as they have been analyzed in Chapter 2, which is why they are particularly useful for this thesis.

First, in their article Häußermann and Lütge (2021) claim that we are currently in the third wave of AI ethics. Looking for principles and guidelines for the ethical use and design of AI was the first wave of AI ethics, while approaches to deal with certain ethical problems, such as the Association for Computing Machinery Conference on Fairness, Accountability and Transparency, formed the second wave. According to the two authors of the article, the first wave played a crucial role in shaping a society in which citizens realize that ethical principles are needed due to the impact that AI has in their lives. The second wave tries to put into practice the guidelines of the first one. Despite the fact that the boundaries of the second and the third waves are not clear, we now live in the third wave and the authors analyze five shortcomings of it. First, AI ethics does not involve the interaction between AI and society and the position that business has in society. Ethics is more focused on the making of AI and does not take as much account of the fact that this technology will then be used by citizens. An AI system may look fair in the laboratory, but ethics should include the implications that this system has when it is used in the "outside world". Second, AI ethics takes for granted that a technical solution is the answer in any problem that may arise. Concerning this, AI ethics takes for granted that AI's progress and development is neutral, but according to Häußermann and Lütge, this is not the case, because people should always have in mind that economic, political and social environment plays a significant role for the development of any technology. They also claim that the second shortcoming implies that AI is the answer to every problem. Before moving to a technical solution, we should wonder if there is another solution aside from improving the AI system. Third, AI ethics focuses on the discrimination, unfairness and biases of an individualist way and it does not see the harm that an AI system may cause to a larger group or to the whole society. Fourth, AI ethics is vague and difficult to implement. Guidelines and ethical principles are mostly impossible to use for the designers of AI due to their theoretical form. Implementing ethical guidelines into AI systems is not an easy task for engineers. Fifth, the relation between ethics and law is undefined. Ethical guidelines may lose their validity when comparing them with laws, but ethics does not have to do with legality. Something that is legal may be unethical, which is why people should separate the two.

To deal with these shortcomings, Häußermann and Lütge (2021) propose a contractualist business ethics; this approach is seldom found in AI ethics. Specifically, as best choice they propose order ethics and compare it with the integrative social contracts theory. According to Häußermann and Lütge (2021, p. 9), "the core elements of order ethics can be summarised as follows:

1. Building on contractualism as normative theory, order ethics argues that ethical conflicts cannot be resolved by reference to overarching normative principles (reasonable pluralism).

2. Instead, ethical conflicts ought to be solved by adapted or new rules to which each stakeholder involved consents based on their individual values.

3. The normative criterion is the mutual advantage that is to be achieved by a respective agreement."

The main difference between order ethics and integrative social contracts theory is that in the latter a macrosocial perspective rules the microsocial perspective. In both cases, the main goal is to deal with problems, despite the fact that in these problems, citizens and corporations with different perspectives and interests may be involved.

The authors suggest and analyze the term "community-in-the loop" using order ethics. In their view, the first question that order ethics should answer is what will happen when there are conflicts due to the different economic and societal positions that citizens and companies may have, concerning many problems that may arise with an AI system. Some concerns about the privacy of a system may be included in the design of the system, but other concerns that users have, may not. The concerns of all stakeholders should be taken into account. As the authors highlight, the participation and deliberation of all stakeholders of AI in the making and use of it is a political task. The values of all stakeholders, including, among others, the designers, users and companies of AI, should be taken into account under equal consideration to ensure pluralism, which is a basic principle of democracy. The heterogeneity of the stakeholders concerning their knowledge in the field of AI and their different experiences and interests should be taken into account in order ethics and should not be an obstacle to achieve their goal.

The steps that according to Häußermann and Lütge (2021, pp. 12-13) should be followed for order ethics in AI are the following. First, everyone should be aware that an AI system may cause conflicts due to the different values of the stakeholders and the different benefits and costs that this system may bring to different groups of people. Second, the community that in any way is related to an AI system should participate in deciding its benefits and its costs. Third, through this deliberation all stakeholders should decide jointly about their ethical rules without competitiveness. Fourth, all stakeholders should agree not on the advantages that may some of them gain, but on the advantages that all of them gain by the implementation of this AI system. Fifth, in this way there will be a pluralistic value creation with the participation of all stakeholders. In this sense, Häußermann and Lütge refer to "human-in-the-loop" (HITL) and "society-in-the-loop" (SITL) to propose "community-in-the-loop" (CITL). HITL is a term referring to the interaction between human and algorithms, while and SITL is a term referring to the interaction between society and algorithms. By "interaction", I mean that human and society are not passive objects that are affected only by AI systems; in fact, they play crucial role through their values, for example in shaping these systems. So, the authors suggest CITL together with order ethics, because these ethics should have a form where

the community will play a determining role in the shaping of AI and AI ethics. In their view, their suggestions for order ethics that come from business ethics, and CITL could tackle the shortcomings that shape the third wave of AI ethics.

A significant work studying the role of designers and society in the making of AI is the article "The ethical AI—paradox: why better technology needs more and not less human responsibility". In this article, De Cremer and Kasparov (2021) claim that AI has no ethics in itself; when we refer to the ethics of AI we refer to the ethics of the people that played a role in the shaping of AI. They believe that people are always responsible for the actions of AI systems. People and not the machines are always accountable. As they claim, big tech supports the idea that technology is always the answer, even for ethical problems. This narrative reinforces the idea that AI machines have their own ethics and decide to act for good or bad. But as De Cremer and Kasparov (2021, p. 2) highlight, "there is nothing magical about AI". It is a technology made by humans and humans are responsible for its actions. As they claim, despite the fact that the majority of society uses phrases such as "the algorithm did it", these phrases "should not even be part of our vocabulary" (2021, p. 2). De Cremer and Kasparov believe that big tech does not actually care about AI ethics, because they always put the progress and development of AI and their profits above the ethical aspects of AI. In their view, big tech companies have created an environment in which AI development always comes first and ethics comes second. They conclude that with this mindset, companies are unable to have true impact in AI ethics for benefiting the whole society.

One of the best ways to understand how big tech deals with the ethics of AI and the role of people who are considered AI ethics specialists in big tech companies, is by studying the article "The Ethical Dilemma at the Heart of Big Tech Companies" by Moss and Metcalf (2019). Moss and Metcalf claim that it is unclear what the stakeholders who are related with AI ethics and work for big tech companies are doing. They refer to these stakeholders as "ethics owners". They interview professionals that work in the ethics departments of big tech and conclude that the difficult part for these professionals is to combine the needs of society together with the needs of tech companies, and to promote AI ethics beneficial for everyone. The authors claim that these professionals receive internal – from their company – and external – from society – pressures for their outcomes. They believe that there are three logics for these pressures. The first is meritocracy, which is the number one value in Silicon Valley. According to it, engineers should evaluate the AI systems and ethics owners walk the line between technical solutions and solutions that come from their knowledge about ethics. Second, market fundamentalism dominates. Due to the fact that big tech companies mostly choose profit over ethical outcomes, ethics owners should have the ability to add value to an AI product when they investigate how to make it more ethical. For example, improving privacy that users will have when using an AI system may avoid future lawsuits against the company producing this system. Third,

technological solutionism is always the answer in Silicon Valley. No matter what the problem is, the solution is always thought to be technical. This is convenient for big tech, but, in the real world, technology is not always the answer. Moss and Metcalf therefore claim that ethics owners have a difficult position in their attempt to avoid these three logics that dominate in big tech. They should aim to include in the design of AI not so much the needs of big tech companies but the needs of society.

Similar concerns about the intentions of big tech companies to develop AI ethics that will truly benefit society are expressed in the article "Big Tech and Data Ethics" by Sam Gilbert. Gilbert (2020) wonders if AI ethics in its current form could actually have a positive impact in people's lives or whether it is just a form of "ethics washing" by big tech. After referring to key themes about data ethics, such as privacy, bias, job displacement, etc., Gilbert claims that virtue ethics is the right form of ethics for AI, because utilitarianism is much more complex in terms of its application to AI and its potential consequences. That is why companies should adopt a policy based on virtue ethics and ask questions such as "What kind of company should we be?" This is the way for them, according to author, to avoid ethics washing and maintain a policy promoting the needs of society.

Finally, Karen Hao (2021), in her article "Big Tech's guide to talking about AI ethics", refers to the attempt of big tech to develop and use a common vocabulary in AI ethics for better communication among tech companies, but also for better communication between them and society. Hao analyzes the meaning of many words and terms that are used by big tech companies and organizations relevant to AI and AI ethics. Having a common vocabulary for everyone is particularly useful for all stakeholders. Some of the terms that Hao analyzes in terms of the way that tech companies use them are the following: *accountability*, which refers to the person who is responsible for the actions of an AI system; *fairness*, which could be defined in many ways, such as the feature of an algorithm making it unbiased; and *transparency*, which refers to the ability of all stakeholders to have access to the data and the code that an AI system uses, etc.

In this section, I analyzed the way that big tech tries to deal with ethics of AI. I mainly based the discussion on articles that in some ways criticize the way that big tech does this. This was useful to establish a more complete picture in the criticism (regardless of perspective) of AI ethics in its current form.

3.2 Presentation of primary research in the media

With all these in mind, it is important to investigate the position that big tech has in our society and how the media presents big tech. To do that, I will use articles from various media outlets but mainly from the *New York Times*. I will investigate articles that have been published in recent years. By

studying the media, I will present the domination of big tech in AI's market and in people's lives, and how citizens are affected by this domination. This approach will allow us to understand the reasons for developing AI ethics by big tech and how these reasons may connect with the design of AI.

In 2014, companies in the IT industry became one of the last to voluntarily disclose their political activities. Silicon Valley in 2014 started to develop a relationship with Washington about Silicon Valley's policy (Willis & Miller, 2014). President Obama, from 2014 to 2016, claimed that big data should be a priority in US policy, as should all the ethical concerns that come with it (Greene et al., 2019). In 2018, the US government invested in AI and announced that AI would be a field for an open debate involving the whole spectrum of aspects (Metz, 2018).

Microsoft, Apple, Amazon, Facebook and Alphabet are the five companies which in 2020 accounted for 20% of the total stock market value (Eavis & Lohr, 2020). Big tech is always expanding and gaining more and more power. And they have just started. They invest huge amounts of money in AI without even knowing where this is going and whether they can manage AI. The five biggest companies mentioned previously earned 166.9$ billion in the third quarter of 2018. As the companies themselves admit, governments do not even know which are the right ways to inspect big tech companies in the field of AI and do not even ask the right questions to control them. As a result, almost everything depends on the companies' values. In a few years, there might not be an aspect of our lives in which big tech companies are not in some way involved (Streitfeld, 2019). Big tech invests most of their money and hopes in cloud storage and software, the most promising fields and the main source of wealth for these companies (Manjoo, 2018).

Figure 7. The logos of popular big tech companies (Source: https://www.publicradiotulsa.org/studiotulsa/2019-11-25/the-growing-role-of-big-tech-in-geopolitics-at-the-tcfr#stream/0)

There are few companies that can afford to pay for data centers and everything else involved in the designing and production of AI. This is why big tech is always getting bigger (Eavis & Lohr, 2020). However, in the last few years, big tech companies have faced a crisis of trust, especially in the field of AI. There are some cases in which federal government trying to inspect the actions of big tech companies, due to the antitrust movement. There are committees from the Justice Department and from the Federal Trade Commission for the investigation of these companies to restore trust in these

companies. Even though citizens should not wait for such committees to make companies behave well, it is a way to control them, and to make them take some measures which might lead to deterring future bad conduct (Condliffe, 2019).

This antitrust movement against big tech arose after many scandals saw the light of day. Both Google and Facebook have a policy of not spreading extremist content and not organizing violence on their platforms. However, both of them sold ads that motivated people to buy merchandises connected with a far-right militia group. This happened despite the fact that the policy of these organizations should not allow such violent groups to be advertised. Moreover, a persistent monitoring tool revealed that Facebook was allowing racial discrimination in advertising houses for sale. Sometimes, such cases have become known as the result of the work of an investigative journalist. Surya Mattu, an engineer who works also as data reporter, believes that tools such as this mentioned about the housing advertisers – which he created – will be vital for people's lives in the future, because persistent monitoring is the only way to control platforms that change all the time (Angwin, 2021).

At the end of 2020, Timnit Gebru, a Google researcher and a co-leader of Google's Ethical A.I. team, said she had been fired due to her criticism on Google's "approach to minority hiring and the biases built into today's artificial intelligence systems" (Metz & Wakabayashi, 2020). She was trying to talk more about the biases of AI and increase the hiring of women and people from minority groups in these fields. After an e-mail she sent to a group that included company's employees, she was fired. Gebru is a well-known scientist in the field of AI ethics and after hiring her, Google painted itself as a big tech company that cared for the ethical aspects of AI. But as it turns out, this is might not be the absolute truth. After a negative answer from Gebru – as she says – to a Google manager, to change some things in an article that she had written about the biases in AI, she was fired. Gebru believes that Google is trying to improve its policy about AI ethics, but without hiring people from all races, sexes, religions etc., this is impossible (Metz & Wakabayashi, 2020).

Activists suggest that citizens should not always trust big tech companies when they say that have a policy on AI ethics, because this might be an easy way for them to get out of the clamor for AI ethics and gain the trust of citizens; what they are actually doing is "ethics washing". That is, they are painting themselves as aware in this field, but in the same time are not doing what is actually needed (Gibney, 2020).

There are six motives for companies to develop a policy about AI ethics (Schiff et al., 2020):

- Social responsibility (the motive to benefit the whole of society and reduce harm)
- Competitive advantage (the motive to gain economic and political advantage against other companies)

- Internal strategic planning (the motive to change the company itself)
- External strategic planning (the motive to intervene to change political, economic and social life)
- To appear as socially responsible (the motive to appear as socially responsible company, no matter whether this is actually true)
- Signal leadership (the motive to be a pioneer in a field and make the difference)

As a result, every company has different standards and different strategies. Governments do not know how to control these big tech companies and at the same time, in most cases the companies do not want to improve their policies by opening the black box of AI, for example, due to competition with the other companies (Murgia & Stacey, 2021). Most of the time, big tech companies give only small hints about how their algorithms work. There are cases where the algorithms are not even controlled by the companies that have made them (Biddle & Zhang, 2021). AI is all over around us and uses our data. Yet only a few companies inform the users about the purposes of their data collection (Biddle & Zhang, 2021). In Silicon Valley, there is the concept of "build it first and ask for forgiveness later" (Singer, 2018). But this is about to change in the logic of engineers in the next few years. In 2018, the Massachusetts Institute of Technology (MIT) and Harvard University started a new course for their students about AI and ethics. At the same time, the University of Texas began to offer a course with the title "Ethical Foundations of Computer Science" (Singer, 2018). In consequence, the ethical aspects of technology and especially of AI might actually become embodied in the culture of an engineer. A company's responsibilities may not have been accurately defined yet, but the MIT concept of an AI which will benefit the whole society is starting to spread all over the scientific community (Markoff, 2016). This is not happening from a suddenly willingness by states and big tech to create an ethically AI, but has arisen through public pressure (Biddle & Zhang, 2021).

As we see, despite the fact that journals and magazines deal extensively with the topic of AI and its connection with big tech and ethics, most of them do not have an STS perspective. An example of an exception is the article "Silicon Valley Pretends That Algorithmic Bias Is Accidental. It's Not" in *Slate* by Amber M. Hamilton (2021). Hamilton (2021) perceives biases that come from AI not as something that came from incorrect calculations and estimations by the engineers, but as something that is part of AI. She claims that we see so many examples of discrimination and bias from AI systems and yet we always believe Silicon Valley's announcements that something went wrong in the design and that after a correction everything will work fine. In a way, Hamilton's view matches better with the view of O'Neil that I analyzed in Chapter 2.3, supporting that the only way to change the nature of AI and its bias is to see tech companies "as institutions that uphold and reinforce structural inequities regardless of good intentions or behaviors of the individuals within those

organizations" (Hamilton, 2021). Hamilton (2021) ends the article by saying, "when we see algorithmic bias as a part of a larger structure, we get to imagine new solutions to the harms caused by algorithms created by tech companies, apply social pressure to force the individuals within these institutions to behave differently, and create a new future in which technology is not inevitable, but is instead equitable and responsive to our social realities."

In this section, I analyzed the way that the media presents big tech and its relation with AI ethics. This is important to realize the interaction between society and big tech, and so the goals that big tech wants to achieve through AI ethics.

3.3 The Microsoft case

This chapter will present research into Microsoft's policy in the field of AI ethics, considering the fact that Microsoft is one of the leading companies in this field. It therefore should be an ideal case study to investigate whether such a company has been affected by STS approaches. Even if this company (which has one of the most adequate policies in AI ethics, as I will explain in this chapter) has not been affected, it means that most of big tech companies have neither. In conclusion, I will recommend AI ethics for Microsoft and for big tech as a whole, by an STS perspective, as it has been analyzed in previous chapters.

First, it must be explained why Microsoft has been chosen over all other big tech companies. In 2014, Microsoft scored 92.9% in an evaluation conducted by activists, companies and associations about its political disclosure. In the same year, Microsoft achieved a position among the top five companies in an annual list for the voluntarily and public disclosure of its political activities (Willis & Miller, 2014). Microsoft is also one of the companies with the most minimalistic and brief principles about AI ethics (Hagendorff, 2020). According to Biddle and Zhang (2021), Microsoft is one of the companies with the highest rating for disclosure of how a user's online content is curated, ranked or recommended. Also, Microsoft is always aware in its systems about the risks of privacy, discrimination and freedom of expression that comes with AI, has joined the partnership on AI[2] and has published a commitment to ethics (Biddle & Zhang, 2021).

The investigation will be done through Microsoft's documents, papers and reports. For example, some papers that I will study and refer to are "Co-Designing Checklists to Understand Organizational

[2] As mentioned in its website "The Partnership on AI (PAI) is a multistakeholder organization that brings together academics, researchers, civil society organizations, companies building and utilizing AI technology, and other groups working to better understand AI's impacts. The Partnership was established to study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society." (Source: https://www.partnershiponai.org/faq/). For more information about PAI: https://www.partnershiponai.org/

Challenges and Opportunities around Fairness in AI" by Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan and Hanna Wallach (2020) (all four are Microsoft researchers), "Responsible bots: 10 guidelines for developers of conversational AI" by Microsoft Corporation (2018), "Datasheets for Datasets" by Gebru et al. (2018), and "Guidelines for Human-AI Interaction" by Amershi et al. (2019).

As researchers of Microsoft suggest (Madaio et al., 2020), many companies have produced AI ethics checklists, in their attempt to guide designers and engineers of AI to make them take the right decisions in the making of AI. But most of the time these checklists are not easy to apply. Engineers may ignore or even misuse them. In many cases engineers do not know how to apply a theoretical checklist for AI. It may be their intention to create a fair AI, but what are the right steps to do it, when fairness has so many aspects? There are also cases in which engineers as employees cannot always do the right thing as they might have wanted to do. They may be afraid of possible dismissal. Moreover, engineers may understand that something has gone wrong with an ethical aspect of their designing, but they do not know how to fix it. Having a checklist gives the impression that a company can guarantee fairness in a company's systems by the fact that its engineers simply follow the rules. But it is not that simple. The policy of a company toward ethical AI should not be limited to a checklist, but should be a complex process including the appreciation of the engineers that the company supports ethical design in all aspects, from the beginning to the end of the construction of AI. Group meetings, seminars, workshops and feedback etc., should synthesize a co-design ethical method for the companies.

The following sentences will present the aspects of fairness checklists as Madaio et al. (2020) perceive them and as Microsoft adopt them. Fairness is not a simple idea, but in fact is a complex concept with many aspects. As a result, giving fairness many definitions ends up with the problem of which of them should be prioritized. Designers do not always have the ability to recognize which of the competing priorities should be chosen. In most of their dilemmas, there are no clear-cut answers. Fairness has sociotechnical aspects. Unfairness may come from societal biases that are embodied in the dataset or from societal factors that come explicitly or implicitly with its designer. Harms that the unfairness of AI could face are many – the people that use AI as well as the people that are represented by it could be harmed directly or indirectly. Harms that come from the unfairness of AI could have different impact in people's lives and seem insignificant. But, in total, many harms that seem insignificant could be extremely harmful for the whole of society. Who decides which of the harms are significant and which not? Finally, we should always have in mind that people who belong to minorities should always be protected and not harmed. It is important to have this in mind because many times these groups are difficult to identify. Many people may belong to multiple minorities. As a result, the solution is not always to consider these group separate (Madaio et al., 2020).

According to Microsoft research, working on AI leads to many failures (Kumar et al., 2019). These failures could be intentional, created by a rival of a company, for example, or by a hacker, or unintentional, when an AI system produces correct outcomes that could be harmful. Machine learning has different characteristics from traditional software, so its failures should be treated as different from what we know so far. A black box architecture gives limited access to an attacker, while on the other hand a "white box" architecture gives the opportunity to attacker to know the AI code and the data. Machine learning systems do not separate failures that come from mistakes from failures that come from intended attacks. In consequence, they must be prepared for all kinds of failure. However, according to Kumar et al. (2019), AI should not transform into a black box over which even the company that has designed it would lose control.

Microsoft (Marshall et al., 2018) supports the idea that AI should not be biased, but at the same time should recognize biases when they happen from people. A good way to achieve this is to expose AI to trolls in a controlled environment. In the same document, Microsoft says it is also carrying out some proposals for certain policy in AI ethics. First, it could do some kind of tests on AI to ensure that AI share human values. Second, training for the democratizing of AI is necessary for the people that design AI. Third, AI algorithms should be in position to recognize and remove data that might be harmful without removing data which is not. Fourth, a central library for establishing the control of AI through transparency and trustworthiness could be created. Fifth, the way that trolls talk could be a useful way for AI to understand what should be avoided. Finally, a framework in which engineers could add the attacks that may happen to an AI system could be created.

Some of the attackers that might appear and that Microsoft is concerned about are cybercriminals with economic or political motivations, insider threats (operating mostly for financial gain) and hacktivists who are political or socially motivated. Some of the attacks may involve social engineering, when users would provide data that in other cases they would not, phishing e-mails, which are based on the weakest link in the security chain, etc. (Kumar et al., 2019).

Microsoft is trying to make transparency one of its main principles. It has documents about its policy in its website that among others refer to the Government Security Program (GSP).[3] Microsoft has created the GSP to achieve transparency through four main pillars. First, Microsoft offers online access for viewing source code through Online Source. Second, Microsoft has five Transparency Centers, in the United States, Belgium, Singapore, Brazil, and China, giving the ability to organizations to inspect and understand the way that Microsoft's AI works. Third, Microsoft through technical data offers information about its cloud services and its products, opportunities to meet Microsoft engineers and understand the way they work, and even visits in Microsoft's facilities for

[3] Source: https://docs.microsoft.com/en-us/security/gsp/programoverview

in-depth conversations. Finally, there is a communication channel about the vulnerabilities and threats that may appear and provides data that may be harmful, called Information Sharing and Exchange.

At the same time, Microsoft believes that to achieve transparency throughout the scientific community it should use a common vocabulary, so the communication between companies and scientists is more efficient (Maclin, 2019). Another important suggestion by Microsoft for transparency in AI bots is that people should always know if they are talking with other humans or with bots. This would ensure that people would know the limitations of their talk (Cheng, 2018).

Microsoft has also 10 guidelines for responsible bots (Microsoft Corporation, 2018):

• The purpose of the bots should be articulated, and should have been decided before even the designing of the bots has begun. Also, there must be metrics for the evaluation of the user's satisfaction about the fulfillment of the bot's purpose.

• The fact that a service is offered by a bot should be known in any case to the consumers. The consumers should know that they are not talking with another person and should have in mind the limitations that a bot may have.

• In any case, there should always be a human available to interact with the consumer if the consumer wants it.

• The bot should be designed so that it does not form any biases against people. The bot should be designed in a way that engineers could always change faults that may appear in its behavior.

• The bot should be reliable. Engineers should build traceability capabilities for any case. Bearing in mind that an AI system may not always provide the correct answer, when a bot is involved in sensitive uses such as delivering services in areas such as health or law enforcement, experts from these fields should always be taken into account when these bots are being designed.

• Bots should treat people fairly. Diversity among designers and engineers is vital for this purpose; the continuous assessment of the data for not being biased is important, too.

• A user's privacy should never be violated by bots. Users should be informed of the extent to which their data will be used, and engineers should collect as little data as they need and not try to get as much as they can. Also, a privacy review and controls for protecting user's privacy are always helpful for the user.

• Engineers should be sure that bots handle data with the appropriate way. This means that bots should be developed in the right way and with correct data.

• Bots should be accessible to everyone. People with disabilities should not be excluded from their use.

• Engineers should be accountable for the bots that they have created. For every action that they made, a human is responsible.

Considering the biases and any other unwanted result of AI that may come from datasets, Microsoft suggests "datasheets for datasets" (Gebru et al., 2018). With datasheets for datasets, Microsoft has the ambition to increase transparency and accountability in the AI field. The company's intention is to improve the needs of two groups related with datasets: the dataset creators and the dataset consumers. If dataset creators work in the right way, then dataset consumers have the ability to use the dataset without unintended biases, faults or unfairness. Some of the information that must be included in the datasheets for datasets are who created the dataset, for what purpose and with whose funding. It should also cover what is included in the dataset, if the dataset has been used previously and if it contains information that may be offensive to other people. Microsoft acknowledges that datasheets for datasets is not the solution to every problem that may arises in this field, but is a start for more transparency and accountability.

Microsoft is giving much attention to the exclusions that may come from AI. That is why it also provides a toolkit to avoid these unintended results. According to Microsoft, designers should always have in mind who they design for. Accessibility should be provided to everyone, no matter the gender, race, abilities, etc. Disabilities are not just health problems but comes together with social norms. AI should be accessible to and inclusive for everyone. The inclusive design may be a chance to increase access, reduce friction and create a more emotional context. The toolkit has three main goals: recognize exclusion; learn from diversity; and solve for one, extend to many (Shum et al., 2016). At the same time, Microsoft has understood the impact that AI has in people's lives, and at the same time the impact that society has in AI's formation. That is why AI has to learn from people and diversity, adapt to the customer's behavior and simultaneously give the opportunity to the customer to adapt (Price & Kim, 2018).

Microsoft also places great emphasis on the biases that emerge from AI. In Microsoft's document "In Pursuit of Inclusive AI", five causes of biases that come from AI are analyzed. First, there is dataset bias, when the biases come from the dataset mainly due to the lack of data with diversity. Second, association bias comes from people that perpetuate unintended biases. Third, automation bias comes from automated decisions that override cultural and social norms. Fourth, interaction bias could be generated when a chat-bot for example interact with people and they convey their own biases, willingly or not. Fifth, confirmation bias comes from the fact that most of the time AI supposes that when a person does something, it means that other people will do the same, something which is not

true because all people are different, and leads to biases against people who make less popular choices, for example (Chou et al., 2018).

Finally, Microsoft has 18 guidelines (Amershi et al., 2019) for designing of AI with the intent to create ethical AI. Microsoft highlights the necessity for the use of these guidelines:

Initially

1. Make clear what the system can do.

2. Make clear how well the system can do what it can do.

During interaction

3. Time services based on context.

4. Show contextually relevant information.

5. Match relevant social norms.

6. Mitigate social biases.

When wrong

7. Support efficient invocation.

8. Support efficient dismissal.

9. Support efficient correction.

10. Scope services when in doubt.

11. Make clear why the system did what it did.

Over time

12. Remember recent interactions.

13. Learn from user behavior.

14. Update and adapt cautiously.

15. Encourage granular feedback.

16. Convey the consequences of user actions.

17. Provide global controls.

18. Notify users about changes.

As we see, these guidelines and the AI ethics of Microsoft in general at some points look like they answer ethical issues that have been analyzed in the AI ethics overview of Chapter 2.2. Even if some guidelines have not been formulated in the same way as the ethical issues presented previously, they may refer to similar values. For example, the first and eleventh guideline refers to transparency. In this point lies the question what *transparency* – and of course all other terms – means for Microsoft and what it means from an STS approach. In conclusion, I will highlight some points concerning the differences between the policy of Microsoft in AI ethics, as presented in this chapter, and a policy from an STS approach.

Figure 8. 6 of the most significant Microsoft's principles for ethical AI (Source: https://cloudblogs.microsoft.com/industry-blog/en-gb/cross-industry/2019/12/11/5-principles-for-ethical-ai/)

4. Conclusion

As I analyzed in the previous chapter, Microsoft has one of the most complete policies in AI ethics among big tech companies. In this thesis, I argue that this policy and this form of AI ethics are not enough; in fact, AI and AI ethics should acquire an STS approach. By an STS approach, I mean investigating the way that AI is designed, the reasons for using it by big tech companies and how this design can play a crucial role in shaping many ethical issues, such as transparency and biases.

Based on the outcomes of Burrell (2016) and O'Neil (2016) that I used in Chapter 2.3 for STS approaches on AI, I will highlight some points that Microsoft could transform or add into its policy about AI ethics for a policy closest to an STS approach.

Two of the most significant aspects of AI ethics are transparency and bias, which most of the time go together. As it has been analyzed, opacity may come from different causes, but the STS approach has

to do with the one that Burrell analyzes most. This is the fact that human mind solves difficult mathematical problems and uses algorithms to do this in different way than machine learning does. As already mentioned in Chapter 2.3, Burrell analyzes two examples to prove this theory. Microsoft does not deal with this aspect of opacity. It does not try to solve the opacity that comes from the fact that, as Burrell (2016) claims, algorithms work at the limits of what the human mind can understand. Microsoft takes it for granted that AI systems are always transparent in their designers and if they are not, something has gone wrong in the design. So, Microsoft with the GSP essentially tries to deal with the other two causes that Burrell refers to as causes of opacity. The four pillars of transparency that Microsoft has aim to deal with the opacity due to the competitiveness among companies and the opacity due to the inability of most people to understand programming languages.

In the previous chapter, I also analyzed "datasheets for datasets" that Microsoft suggests for achieving transparency. This is also not an STS approach due to the fact that this approach invests in better design by using better and unbiased data. But as O'Neil (2016) explains, there are many cases in AI where designers did not want to use biased data or had in mind using the right data for an AI system, but in the end the output of the system was in some cases biased. As a result, an STS approach in the example of "datasheets for datasets" should have in mind that an AI system could lead to biased outcomes even if designers think that they have used unbiased data. This is because data may be biased in a different way from usual, and designers could not perceive it or data may become biased after the processing and the way they are used by the algorithms. Data that may not be biased at a primary stage may become so when used in a certain way. In the case that designers cannot understand the way that an AI system take decisions, they might not be in a position to understand if the data is biased at one point of the process.

One of the most significant terms when STS approaches are referred to is the "black box". Microsoft rarely refers to AI as a "black box" and never in reference to an STS approach. There is a huge difference between the opacity of AI that Microsoft tries to deal with by exposing its designing to the governments and to the consumers and "black box" of AI that O'Neil in her book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* analyzes. Testing AI systems, and training AI engineers in how to develop fair AI, which I analyzed in Chapter 3.3, are some of the ways that Microsoft uses to deal with biases. O'Neil, on the other hand, perceives AI as a "black box" which generates and reproduces biases not because AI engineers are not trained enough. She suggests that we see AI as a "Weapon of Math Destruction". Here lies the STS perspective. AI is not a "black box" only to those who do not know how to program; in many cases, it is a black box to AI engineers, too. This is the outcome by the article of Burrell, and as O'Neil explains, companies take advantage of this opacity to dress the results of AI systems as neutral, fair and right. O'Neil explains with many examples that tech companies use the black box of AI systems for their own

benefit. That is why she uses the term "Weapons of Math Destruction" for AI systems. Microsoft should include this aspect if it wants to reduce opacity and biases. Microsoft guidelines for the designing of AI referring, for example, to defining the purpose of an AI system or to understanding what is going wrong in an AI system are not enough. The STS perspective has to do with the concerns that exist even if AI systems look like they have been designed with the "right way". These concerns come from the fact that the "right way" comes together with the analysis of Burrell about the designing of AI that generate the "black box", and the analysis of O'Neil in taking advantage of "black box" of AI to use it as a "Weapon of Math Destruction".

I therefore carried out a historical contextualization of AI and referred to the broken promises, empty threats and winters of AI, as Garvey highlights. I did this to give an STS perspective in AI concerning that the development of AI is affected by social, economic and political environment prevailing at the time. Second, I did an overview in AI ethics and analyzed all the ethical issues that come with AI such as privacy, biases, employment, etc. This overview is necessary to understand AI ethical issues in their current form and so to investigate if they are adequate after comparing them with the STS approach of AI ethics. Third, I referred to O'Neil's book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* and Burrell's article "How the machine 'thinks': Understanding opacity in machine learning algorithms" to give an STS approach in AI and AI ethics, and more specifically to investigate transparency, black box and biases from an STS perspective. Fourth, I studied the role of big tech in determining AI and AI ethics. According to the secondary literature, there are many criticisms in AI ethics that big tech has developed, and have various perspectives. One of the criticisms is that big tech companies give priority to their profits and not to actual ethics that will benefit society. At the same time, big tech companies, as Gilbert (2020) suggests, are doing "ethics washing" in their try to avoid criticism from society and they do not actually develop AI ethics by a perspective which will benefit society as an STS perspective in AI ethics. The articles featured in Chapter 3.1 were necessary to develop a complete picture about AI ethics and to see if there are criticisms close to STS approaches. At the same time, studying the media gave me the opportunity to present the domination of big tech companies in the field of AI and the consequences that this domination has in AI ethics. As Hamilton (2021) indicates, citizens should not look only the small picture and wait for big tech companies to technically correct their mistakes that, as they claim, are responsible for unintended biases, but in fact a first step to eliminate biases in AI is for citizens to look the broader picture and understand the role of big tech and AI in society. Finally, I analyzed the policy of Microsoft, one of the most adequate policies by a big tech company in AI ethics, and concluded that despite the fact that Microsoft has tried to deal with all ethical issues that have arisen in the field of AI, something is missing.

I suggest that a totally adequate and complete policy in AI ethics should involve an STS perspective, as I described it above. This is the only way for AI to truly benefit society. This is the only way to talk about real AI ethics.

Bibliography

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*, *3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Chou, J., Ibars, R., & Murillo, O. (2018). *In Pursuit of Inclusive AI*.

Coeckelbergh, M. (2020). *AI Ethics*. The MIT Press.

De Cremer, D., & Kasparov, G. (2021). The ethical AI—paradox: why better technology needs more and not less human responsibility. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00075-y

Garvey, C. (2018). Broken Promises and Empty Threats: The Evolution of AI in the USA, 1956-1996. *Technology's Stories*, *6*(1). https://doi.org/10.15763/jou.ts.2018.03.16.02

Garvey, S. C. (2021). Unsavory medicine for technological civilization: Introducing 'Artificial Intelligence & its Discontents.' *Interdisciplinary Science Reviews*, *46*(1–2), 1–18. https://doi.org/10.1080/03080188.2020.1840820

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2018). *Datasheets for datasets*. http://arxiv.org/abs/1803.09010

Gibney, E. (2020). The Battle To Embed Ethics in AI Research. *Nature*, *577*. https://doi.org/10.1038/d41586-020-00160-y

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2122–2131. https://doi.org/10.24251/HICSS.2019.258

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*,

30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Häußermann, J. J., & Lütge, C. (2021). Community-in-the-loop: towards pluralistic value creation in AI, or—why AI needs business ethics. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00047-2

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *2020 CHI Conference on Human Factors in Computing Systems*. http://dx.doi.org/10.1145/3313831.3376445

Microsoft Corporation. (2018a). *Responsible bots: 10 guidelines for developers of conversational AI*.

Microsoft Corporation. (2018b). *The Future Computed: Artificial Intelligence and its role in society*. Microsoft Corporation.

Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. In *The Stanford Encyclopedia of Philosophy* (Winter 2020). https://plato.stanford.edu/archives/win2020/entries/ethics-ai/

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (1st ed.). Crown.

Price, M., & Kim, D. (2018). *Respecting Focus: A Behavior Guide for Intelligent Systems*.

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data and Society*, 1–5. https://doi.org/10.1177/2053951720942541

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's next for AI ethics, policy, and governance? A global overview. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–158. https://doi.org/10.1145/3375627.3375804

Shum, A., Holmes, K., Woolery, K., Price, M., Kim, D., Dvorkina, E., Dietrich-Muller, D., Kile, N., Morris, S., Chou, J., & Malekzadeh, S. (2016). *Inclusive: Microsoft Design*.

Internet Material

Angwin, J. (2021). *Big Tech Needs Persistent Monitoring*. https://www.getrevue.co/profile/themarkup/issues/big-tech-needs-persistent-monitoring-308469

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Berg, N. (2014). Predicting crime, LAPD-style. *The Guardian*. https://www.theguardian.com/cities/2014/jun/25/predicting-crime-lapd-los-angeles-police-data-analysis-algorithm-minority-report

Biddle, E. R., & Zhang, J. (2021). *Moving fast and breaking us all: Big Tech's unaccountable algorithms*. Ranking Digital Rights. https://rankingdigitalrights.org/index2020/spotlights/unaccountable-algorithms

Cheng, L. (2018). *Microsoft introduces guidelines for developing responsible conversational AI*. https://blogs.microsoft.com/blog/2018/11/14/microsoft-introduces-guidelines-for-developing-responsible-conversational-ai/

Condliffe, J. (2019). The Week in Tech: What Not to Expect From Big Tech's Antitrust Showdown. *The New York Times*. https://www.nytimes.com/2019/06/07/technology/big-tech-antitrust.html?searchResultPosition=3

Eavis, P., & Lohr, S. (2020). Big Tech's Domination of Business Reaches New Heights. *The New York Times*. https://www.nytimes.com/2020/08/19/technology/big-tech-business-domination.html?searchResultPosition=2

Gilbert, S. (2020). *Big Tech and Data Ethics*. Bennett Institute for Public Policy at the University of Cambridge. https://www.bennettinstitute.cam.ac.uk/blog/big-tech-and-data-ethics/

Hamilton, A. M. (2021). Silicon Valley Pretends That Algorithmic Bias Is Accidental. It's Not. *Slate*. https://slate.com/technology/2021/07/silicon-valley-algorithmic-bias-structural-racism.html

Hao, K. (2021). Big Tech's guide to talking about AI ethics. *MIT Technology Review*. https://www.technologyreview.com/2021/04/13/1022568/big-tech-ai-ethics-guide/

Kumar, R. S. S., O'Brien, D., Albert, K., Viljöen, S., & Snover, J. (2019). *Failure Modes in Machine Learning Systems*. https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

Maclin, V. (2019). *Solving the challenge of securing AI and machine learning systems*. https://blogs.microsoft.com/on-the-issues/2019/12/06/ai-machine-learning-security/

Manjoo, F. (2018). Stumbles? What Stumbles? Big Tech Is as Strong as Ever. *The New York Times*. https://www.nytimes.com/2018/08/01/technology/big-tech-earnings-stumbles.html?searchResultPosition=4

Markoff, J. (2016). How Tech Giants Are Devising Real Ethics for Artificial Intelligence. *The New York Times*. https://www.nytimes.com/2016/09/02/technology/artificial-intelligence-ethics.html?searchResultPosition=4

Marshall, A., Rojas, R., Stokes, J., & Brinkman, D. (2018). *Securing the Future of Artificial Intelligence and Machine Learning at Microsoft*. Microsoft. https://docs.microsoft.com/en-us/security/engineering/securing-artificial-intelligence-machine-learning

Metz, C. (2018). Artificial Intelligence Is Now a Pentagon Priority. Will Silicon Valley Help? *The New York Times*. https://www.nytimes.com/2018/08/26/technology/pentagon-artificial-intelligence.html?searchResultPosition=1

Metz, C., & Wakabayashi, D. (2020). Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I. *The New York Times*. https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html?smid=tw-share

Moss, E., & Metcalf, J. (2019). The Ethical Dilemma at the Heart of Big Tech Companies. *Harvard Business Review*. https://hbr.org/2019/11/the-ethical-dilemma-at-the-heart-of-big-tech-companies

Murgia, M., & Stacey, K. (2021). *How is Big Tech dealing with ethical problems?* Financial Times. https://www.ft.com/content/0c91c30e-f3c6-4219-bd33-475cc721ffbe

Singer, N. (2018). Tech's Ethical 'Dark Side': Harvard, Stanford and Others Want to Address It. *The New York Times*. https://www.nytimes.com/2018/02/12/business/computer-science-ethics-courses.html?searchResultPosition=2

Starr, S. B. (2014). Sentencing, by the Numbers. *The New York Times*. https://www.nytimes.com/2014/08/11/opinion/sentencing-by-the-numbers.html

Streitfeld, D. (2019). Big Tech May Look Troubled, but It's Just Getting Started. *The New York Times*. https://www.nytimes.com/2019/01/01/technology/big-tech-troubled-just-getting-started.html?searchResultPosition=3

Turque, B. (2011). 206 low-performing D.C. teachers fired. *The Washington Post*. https://www.washingtonpost.com/local/education/206-low-performing-dc-teachers-fired/2011/07/15/gIQANEj5GI_story.html

Willis, D., & Miller, C. C. (2014). Tech Firms and Lobbyists: Now Intertwined, but Not Eager to Reveal It. *The New York Times*. https://www.nytimes.com/2014/09/25/upshot/tech-firms-and-lobbyists-now-intertwined-but-not-eager-to-reveal-it.html?_r=0