AI, Ain't I a Woman? Seeking a New Feminist Ethics of Technology, Between Algorithmic Decision-Making Processes and the European Legislation

A Thesis

Submitted to the Faculty of Humanities

University of Leiden

In Partial Fulfilment of the Requirements

For the Degree of

Master of Arts

in Philosophical Perspectives on Politics and the Economy

by

Diletta Huyskes

student number 2603993

Word count: 18.183

July 2020

Supervisor: Dr. Jan Sleutels

Second Reader: James McAllister

TABLE OF CONTENTS

«My heart smiles as I bask in their legacies
Knowing their lives have altered many destinies
In her eyes, I see my mother's poise I
n her face, I glimpse my auntie's grace
In this case of deja vu
A 19th century question comes into view
In a time, when Sojourner truth asked
"Ain't I a woman?"

Today, we pose this question to new powers
Making bets on artificial intelligence, hope towers
The Amazonians peek through
Windows blocking Deep Blues
As Faces increment scars
Old burns, new urns
Collecting data chronicling our past
Often forgetting to deal with
Gender race and class, again I ask
"Ain't I a Woman?"

Face by face the answers seem uncertain
Young and old, proud icons are dismissed
Can machines ever see my queens as I view them?
Can machines ever see our grandmothers as we knew them?
Ida B. Wells, data science pioneer
Hanging facts, stacking stats on the lynching of humanity
Teaching truths hidden in data
Each entry and omission, a person worthy of respect

Shirley Chisholm, unbought and unbossed
The first black congresswoman
But not the first to be misunderstood by machines
Well-versed in data drive mistakes

Michelle Obama, unabashed and unafraid
To wear her crown of history
Yet her crown seems a mystery
To systems unsure of her hair
A wig, a bouffant, a toupee?
May be not
Are there no words for our braids and our locks?

Does sunny skin and relaxed hair
Make Oprah the first lady?
Even for her face well-known
Some algorithms fault her
Echoing sentiments that strong women are men
We laugh celebrating the successes
Of our sisters with Serena smiles
No label is worthy of our beauty.»[1]

[1] www.notflawless.ai Poet of Code shares "AI, Ain't I A Woman" - a spoken word piece that highlights the ways in which artificial intelligence can misinterpret the images of iconic black women: Oprah, Serena Williams,

INTRODUCTION

Automated decision-making processes (ADMs) represent the latest contemporary paradigm of technological evolution in society. In brief, these are highly elaborated systems belonging to the broader category of Artificial Intelligence (AI) technologies. Technologies like these are becoming increasingly dominant in our lives. AI technologies, often powered by algorithms of varying sophistication, have become responsible for giving us access to mortgages, credit scoring, job positions, pensions, state aid, and can have decisive effects on our lives by providing medical treatment, predictions of recidivism, and even arrests. We are talking about the same systems at the base of any calculation system and the same algorithm that determines as spam some emails that arrive in our mail folder. This corresponds, however, to their most advanced evolution, which in many cases coincides with a fully automated process resting on *machine learning* (ML) algorithms that no longer need human intervention to make decisions. My choice to address algorithmic decisions in particular and not AI technologies in general aims to draw attention to these systems, introduced to delegate - and disempower - decisions that were previously made by human beings.

The implications of the use of automated algorithms to every component of daily life have been researched by many authors in the field of ethics of technology, philosophy, sociology and politics. This area of research presents a continuous evolution, since the speed of technological innovation leads to increasingly rapid and frequent changes. The most interesting – and at the same time worrying – matter for investigation concerns the prejudicial component of these systems. Namely, *algorithmic bias* refers to the output of a decision that disadvantages certain individuals or social groups, causing repercussions on many fronts of their existence. It has been demonstrated how systematic these biases are and how susceptible these systems are to these risks. As cases grew, therefore, the urgency of addressing this issue in the public debate became apparent. Discrimination usually occurs against categories that are already socially excluded, giving rise to race, gender or class biases. Each algorithm, in fact, is trained on data sets collected from different social contexts, and which can later give rise to unfair combinations for many different reasons. Nevertheless, the laws and policies that revolve around the use of AI are still at an embryonic stage of development, and where enforced they can still manifest problems of straightforwardness and applicability.

In this dissertation, I decided to focus on the relation between ADMs and gender bias. This is because, despite the validity and importance of other studies on bias, the research and

Michelle Obama, Sojourner Truth, Ida B. Wells, and Shirley Chisholm. https://www.youtube.com/watch?v=QxuyfWoVV98 attention on this issue is particularly outdated. This thesis is built to reach a gradual achievement of the final objective, i.e. the proposal to structure a new feminist ethics of technology that responds to the challenges of the present. In order to do this, I have assigned the first chapter a more technical role, to explain in detail what automated decision-making processes are and why they should interest us. Through the explanation of their mathematical functioning, we can already get an idea of their opacity and complexity, which then leads to the incorporation of prejudices and the creation of new stereotypes. This also requires to address the problem of the «black boxes», a highly analyzed phenomenon that indicates the inscrutability of what happens between the input and the output of an algorithm, making it very difficult to unveil the decision that has been taken. The second part of the first chapter is dedicated to providing an analysis of the political and legislative framework in the European context, in order to clarify any doubts about the measures adopted to regulate the use and the impact of automated decision-making processes. The attention to the European context is motivated by the desire to demonstrate that, although it is commonly believed that this territory guarantees more and more powerful protections against the use of these technological systems, we are subject to the same risks than other countries, because we still share the same mentality of unconditional and unquestioned trust in technological neutrality. In particular, I will demonstrate how the relevant European legislation, the General Data Protection Regulation (GDPR), if analyzed in detail, presents some deficiencies that are difficult to justify. It is widely believed that the GDPR provides the citizens subject to automated decisions with a «right to explanation» (Goodman and Flaxman, 2016). If implemented, such a right would allow to receive important information about discrimination, where it occurs, and why it happened. Nevertheless, I will support the idea advanced by Wachter et al (2017) to show that a valid right to explanation cannot be found within the GDPR, together with a serious intention to implement such a right, and provide reasons for my choice.

In the second chapter, I will introduce the problem of gender bias, i.e. the discrimination against women that is exacerbated and intensified through decisions made by machine learning systems. In order to support my thesis and demonstrate my claims, this section will provide many empirical cases of women discrimination carried out by ADMs, presented by different authors and studies. Although it may appear to be a minor problem, «technological design is particularly important, as it often captures and reproduces controlling and restrictive conceptions of gender which are then repetitively reinforced» (Collett and Dillon 2019, 4). Importantly, I will distinguish between several types of algorithmic biases, to collocate them within the complex technical process and show how they originate. What will emerge is that the role and ideas of the designers and actors involved in the realization of these technologies have a particular impact on the final output. In this context, the concept of «data representativeness» will be particularly relevant to understand how the lack of female representation favors the emergence of biases. In this chapter, then, I will point out that every algorithmic error has its origin in a corresponding social, and therefore human, prejudice, introducing Andrew Feenberg's concept of «technological code» and the idea of an alleged technological neutrality, that together with determinism plays an important role in unveiling the historical causes of biases.

Finally, the third chapter will present the philosophical and ethical framework in which this debate takes place: the introduction of the idea of technology as a political phenomenon. This consideration was introduced as a starting point by social constructivists, whose research was collocated in the more general area of science and technology studies (STS) and recognised the need to bring technology back in an political and empirical perspective. Historically, technological tools are attributed to economic achievements and profit maximization, which have slowly moved them away from experience. This has led us to think of technology as a metaphysical force beyond our control, turning any kind of human intervention over it *de facto* futile. These issues are bound to a normative and philosophical domain, which some philosophers have identified in the thesis of neutrality. This corresponds to the logic behind the incontestability of automated decision-making systems and the generic nature outlined by the GDPR.

The second part of the chapter aims to introduce feminist studies of technology (FTS), born to integrate the debate with a specific focus on gender and on different social groups that classic constructivism has failed to recognize. Feminist studies on gender and technology aimed at revealing the patriarchal nature of technology and its relevance for any study that seeks to analytically examine the social construction of technology, claiming that «since technology and gender are both socially constructed and socially pervasive, we can never fully understand one without also understanding the other» (Lohan and Faulkner 2004, 319). I will draw upon the works of leading authors such as Judy Wajcman, Cynthia Cockburn and Donna Haraway to highlight their contribution to feminist theories in the 1990s and early 2000s. In order to analyze the relationship between automated decision-making processes, highly advanced forms of AI and gender bias, it is necessary to introduce new theoretical approaches, which are necessary due to the continuous technological transformation. This thesis ultimately seeks to provide guidance and encourage a new feminist ethics of technology, which should address any aspect analyzed in the course of this research and move beyond it, establishing a situated practice of investigation that does not fall in the same error of neutrality. Conversely, it should require a new framework for legislators and policy makers that clearly directs algorithmic discrimination, and most importantly it should demand new «explanations» that do not hide behind technological neutrality.

1. Automated Decision Making

Automated decision-making (ADM) is «the ability to make decisions by technological means without human involvement» (Working Party Guidelines, Art. 29, 2018) which belongs to the broader area of Artificial Intelligence (AI) technologies. The latest are intended as machines or computers that imitate cognitive functions typical of the human mind, such as learning and problem solving, but also translation, driving or recommending a book. The development of these technologies was possible because of a general optimism towards the idea that human cognition is easily replicable, especially when thinking about the greater precision ensured by automated systems, given by the identification of statistical links that always operate according to the same set of rules. For this reason, these technologies are considered particularly useful when used to maximize profits.

Usually, AI technologies revolve around the use of algorithms. Despite the lack of a general and shared definition in the academic debate, an algorithm is a sequence of instructions that a mechanical computer can execute, designed to complete a task or solve a problem. Robin K. Hill defines an algorithm «as a mathematical construct with a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions» (Hill 2016, 47). Algorithms can be implemented for several reasons, however the interest of this thesis focuses on their application in decision-making processes. Almost in all cases, algorithms implemented for decision-making and predictive purposes are based on *machine learning* techniques, i.e. automated improvement through experience. Machine learning is «any methodology and set of techniques that can employ data to come up with novel patterns and knowledge, and generate models that can be used for effective predictions about the data» (Van Otterlo 2013, 46). As a consequence, unlike other programs, machine learning systems do not necessarily need explicit human rules in order to achieve a goal. The machine derives its decisions on the basis of the data and algorithms on which it has been trained, learning autonomously from the correlations and patterns it identifies in data recurrences. This practice is also known as *pattern recognition*.

It is easily conceivable that the self-learning method typical of these systems poses some ethical concerns, where the algorithm's own decision was previously carried out by a human being. Once we understand the immense territory in which these technologies operate, it is hard to imagine our lives today to be exempted from some kind of automated decision-making. This kind of processes, closely related to the collection of personal data, probably represent the technological transformation that most pervades the daily life of humans in this historical moment. Their power, in fact, is that of being used as a means of organizing their social, bureaucratic and political existence and even satisfying their political needs. The resulting consequences of the use of machine learning techniques produce a continuous impact on society.

Even arguing that algorithms should not replace decisions made by human beings, it is undeniable that this is already happening, regardless of any ethical concern. To analyze the algorithmic impact on society and the sometimes extreme repercussions resulting from it, it is essential to look at their technical and mathematical functioning. What we need is a deep discernment of the rules dictated by computer science, by its supposed undeniable determination, in order to question its real-life connections. To understand the new frontiers of power, it is necessary that philosophy fully understands how they operate. For the purpose of analyzing utopia, and criticize it, it is necessary to learn how to use its own language. Only once this has been done can we deepen the links between technique and progress, between logic and value production, and address its long history of concerns.

1.1 Algorithmic Relevance and Functioning

«The goal of a learning algorithm is to build a function, or *classifier*, that assigns a class label (e.g. 'spam') to any object (e.g. 'emails') that has not yet been labelled» (Scantamburlo et al 2018, 15). In other words, an algorithm is always implemented to reach a conclusion and have a certain type of effect. In order to make certain decisions, the algorithm is trained with a series of data that are useful for the purpose of the task. As Mittelstadt et al. underline, besides being a mathematical construct, the algorithm is a configuration, an implementation, an artifact. The most relevant public use of machine learning algorithms is «to make decisions, e.g. the best action in a given situation» (Mittelstadt et al 2016, 2).

The principle behind the use of automated decisions is usually that outcomes, in the form of scores and/or results, can be used as indicators of risk and opportunity, especially for future behaviors: «depending on the score produced, the algorithm triggers a certain response action such as 'detaining an offender' or 'rejecting a loan application'» (Scantamburlo et al 2018, 3). According to Lilian Mitrou, machine learning has a close link with prediction, since the task of these models is to link past behaviors to outcomes that can predict the future. These are used to decide on factors that are crucial to the lives of people, in financing, working, living and other spheres (2018, 13).

Generally, the work of a machine learning algorithm is multiple, and operates in two parallel ways: a *classifier* and a *learner*. The classifiers take the input (defined as a set of *characteristics*) and produce an output (a *category*). A classifier, therefore, takes a set of characteristics and produces a decision, or output, choosing between different categories. As Jenna Burrell explained, for example, an automated decision making system dealing with disease diagnosis can «take input (clinical presentation/symptoms, blood test results) and produce a disease diagnosis as output ('hypertension', 'heart disease', 'liver cancer'). However, machine learning algorithms called learners must first train themselves on the test data. The result of this training will then be used by the classifier to determine the classification for the new input data, which may for example correspond to historically previous data» (2016, 5). In principle, the model can be taught to the algorithm through human labeled inputs and supervision that will then include human presence. However, in other cases, the algorithm acts alone, defining the most suitable models to make sense of a set of inputs, with no need to understand the underlying causal mechanism that generated those data. In most cases, indeed, the designers of an artificial intelligence system decide the characteristics that serve to describe an object and the classes or categories available, providing the inputs but not designing the actual decision function. In machine learning, the decision model is automatically generated using many training examples and labeled data. The fundamental choice of the designer is to determine the category or class of possible decision functions from which the system can then choose to reach the goal. Unfortunately, it is not an essential requisite that the human controller understands the logic behind the decision-making process, as the algorithm is intended to operate independently (Mittelstadt et al, 2016, 3).

As Mittelstadt et al suggest, «causality is not established prior to acting upon the evidence produced by the algorithm» (2016, 5). The search for causal links is difficult, and I will later show how correlations established in such huge datasets are frequently very difficult to explore in depth. The increasing use of ADM systems by private agents, governments, state organizations and public administrations is particularly useful when it significantly lightens human workload. The use of a machine to analyze data and identify a satisfactory response based on a certain input reduces the need for human labor. However, according to technicians, their use guarantees greater accuracy and efficiency in identifying similar patterns. In many countries, with a special mention to the United States, these systems are implemented in several areas of high decision-making responsibility, such as criminal justice, law enforcement, recruitment decisions, credit scoring, school allocation mechanisms, health care and assessment of eligibility for public benefits.

The most important component of any algorithmic decision-making process, especially when fully automated, are our data, used from the very beginning for the learning process. This is the reason why they have become such a valuable commodity in modern times. As Shoshana Zuboff defines it, we are living in the age of *surveillance capitalism*, where our data are «at stake in a new economic order that claims human experience as free raw material for translation into behavioral data» (2018, 8). We shall notice a power struggle between the commercial interest of who owns the data and data subjects[2] who claim for transparency. While the means of production in the data economy are subordinated to our emotions and transform the utility that our profiles render to the market, there is too little information about these systems and how they affect our lives. According to most of the ethical theories built around technology, indeed, algorithms are extremely value-laden, in the sense that the choices they reach are seemingly innocuous, but they actually have the power to change people's lives. Operating parameters are entered by developers and configured by users, often unconsciously. The resulting outputs exhibit some dominant values, or at least preferences over some values rather than others. This operation, however, does not guarantee that the conduct carried out by the algorithm is ethically acceptable. In the next chapter, I shall demonstrate how algorithms may show discriminatory behavior due to different programming processes.

1.2 Data Sources and Black Boxes

As examples of new technologies, ADMs and algorithms inherit an ethical challenge that includes the collection, decryption and organization of huge amounts of data. This process may include profiling, that is «any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyze or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behavior, location or movements» (GDPR 2018).

The types of personal data that can be used for profiling or decision-making are very different: personal, judicial, regarding our communications to third parties, our geo-location, but especially those that reveal our ethnicity, religious beliefs, philosophical and political opinions, union membership, our health and relations. Among these are biometric and genetic data, useful for facial, vocal, postural or emotional recognition systems. According to the UK's Information Commissioner Office (ICO), «organizations obtain personal information about individuals from a variety of different sources. Internet searches, buying habits, lifestyle and behavior data gathered from mobile phones, social networks, video surveillance systems and the Internet of Things are examples of the types of data organizations might collect. They analyze this information to classify people into different groups or sectors. This analysis identifies correlations between different behaviours and characteristics to create profiles for individuals. This profile will be new personal data about that individual» (ICO, Guide to the GDPR, 2018).

[2] According to Article 4: Definitions GDPR, data subject is the natural person to whom data relates.

Far from being a new phenomenon, Bowker and Star demonstrate that in a classification, «each category values one point of view and silences another». The consequences can go so far as to make the lives of individuals «broken, twisted and tormented by their encounters with classification systems» (Bowker and Star, 1999). Although this represents a common and widespread process, it is clear that in recent years data extraction has become much more pervasive. Indeed, «since the accuracy of these algorithms is known to improve with greater quantities of data to train on, the growing availability of such data in recent years has brought renewed interest to these algorithms» (Burrell 2016, 5). The collection of personal data useful for all these purposes shall happen in line with what is indicated and limited by the General Data Protection Regulation (GDPR)[3], the latest European regulation on the use of data, which will be addressed in detail hereafter.

«When data are used as (or processed to produce) evidence for a conclusion, it is reasonable to expect that the connection between the data and the conclusion should be accessible (i.e. intelligible as well as open to scrutiny and perhaps even critique)» (Mittelstadt et al 2016, 4). In this case, the connective passage would be intelligible if it provided the rationale behind the decision. The decision-making process, before reaching a conclusion, both in humans and machines could potentially provide a lot of information about the conclusion that is reached. In the case of human beings, this is a complicated process, but certainly approachable in different ways, first of all dialogue. Accessing the processes performed by machine learning algorithms, however, can be highly complicated. As Mittelstadt et al claim, «the rationale of an algorithm can be incomprehensible to humans, rendering the legitimacy of decisions difficult to challenge. Besides being accessible, information must be comprehensible to be considered transparent» (Mittelstadt et al 2016, 7). We are talking about systems that rely on neural networks, i.e. extremely complex processes modeled on the human brain and «based on millions of computing clusters. The software constantly updates and changes its nodes in response to new data, resulting in extremely complex codes and calculations» (Silva and Kenney 2018, 23). For these reasons, certain AI technologies are defined as «black boxes». This complexity is highly problematic because the algorithms can be biased, potentially leading to discriminatory results. This problem, in fact, does not remain anchored to the technical field of technology, but has important repercussions in society.

The «black box» metaphor has been introduced by studies on cybernetics. In this context, it refers to a system of which we can only know the inputs and outputs, but not the central process. It means that the machine cannot explain the reasons behind a specific decision that

[3] Reg (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Dir 95/46/EC (General Data Protection Regulation) 2016.

itself has taken: the algorithm has an impact on reality, changing the state of things, but we cannot know why. According to Jenna Burrell, algorithms «are opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs» (Burrell 2016, 1). «Both the inputs (data about humans) and outputs (classifications) can be unknown and unknowable.» (Ibid.).

Burrell notes that there are many factors that contribute to the inscrutability of the machine, that often conflate. One is that of proprietary concerns (or corporate secrecy): the inaccessibility to algorithms is justified by a competitive advantage, in order not to reveal the technological means used and to be ahead of commercial opponents. (Ibid., 3). In this case, the general functioning and the computational design are often accessible. Nevertheless, as far as the transparency of the algorithm is concerned, we shall discuss later on how, also from a legal point of view, it is not acceptable for accessibility to be traced back to a mere vision of machine's general rules. Frank Pasquale defines opacity as a «remediable incomprehensibility» (2015, 7) which could result from the willingness of companies to protect themselves for competitive reasons, but also from the need to cover up certain discriminatory intentions or to evade certain norms (Burrell, 2016, 4). For this type of opacity, one possible solution could be to make the algorithm's internal code available for scrutiny, through regulatory means. A sort of open source would be possible if companies were willing to open up their algorithm design, which is unlikely. Therefore, in order to protect their design, Pasquale proposed «the use of an independent auditor who can maintain secrecy while serving the public interest» (2015, 141).

Besides competitive advantage, other reasons for the algorithms to be inscrutable are national security, privacy, or specific legal issues such as trade secrets. In some cases, different forms of opacity are combined. Most AI algorithms are proprietary and have commercial value, like the ones used by Google search and Facebook news feed, thus protected by trade secrets (Noto La Diega 2018, § 34). In the same research, he explains why also individual property rights play an important role and are very difficult to open. One last black box, not less important, is the technical one: analyzing and studying the software that implements the algorithms requires high computational skills – certainly not belonging to the general public – hence the «need to ask an expert third party to carry out such activities on behalf of the lawful user of the software» (Noto La Diega, 2018, § 42). «Opacity in machine learning algorithms», then, is also «a product of the high dimensionality of data, complex code and changeable decision-making logic» (Burrell, 2016, 1). Specifically, algorithmic illiteracy is also a huge issue. Not necessarily a form of opacity, technological illiteracy is a widespread phenomenon because of the lack of a basic education on the existence of these systems. As machine learning researcher Pedro Domingos points out, for example, to use technology effectively it is not vital to understand every little detail of its internal functioning. Instead, it is necessary to have a good conceptual model of it. We need to be able to understand the algorithms in a simplified way so as not to overestimate or underestimate their powers (Domingos, 2016, 44).

As Burrell concludes, in order to address the issues posed by the black box and thus ensure greater transparency, a multilateral approach is needed, through multiple checks on the algorithm and its codes, but also by raising awareness among designers and educating the general public. Another alternative is the use of tools such as open source. (2016, 10). Machine learning algorithms do not identify causal effects (and are not designed to do so). They can only represent probabilistic associations, so they do not constitute sufficient and necessary tools to deal with real-world situations. As Scantamburlo et al explained, their «predictions or classifications are educated guesses or bets, based on large amounts of data, and can be expected to work subject to certain assumptions» (2019, 7). Understanding causal relationships, however, is the only way to evaluate the impact of an intervention on reality, as causality allows us to reason in terms of «counterfactuals» (i.e. what would have happened in an alternative scenario) (Angrist and Pischke, 2008). In order to finally understand how to make sense of all this difficulties that revolve around algorithmic decision-making and its transparency, we shall have a look at the European framework of laws that regulates it, to judge if it is an adequate and sufficient tool to tackle the problems that will be specifically presented in the next chapters.

1.3 ADMs in Europe: the GDPR

As mentioned before, algorithmic decision-making is addressed in Europe under the data protection authorities – by the General Data Protection Regulation (GDPR). The Regulation was introduced in 2016 to replace the 1995 Data Protection Directive, which was adopted at a time when the Internet was at a much less advanced stage and the resulting risks were more limited. The decision to introduce the GDPR (EU Regulation 2016/679) came at a time when subjects' personal data were becoming increasingly popular and valuable. Above all, to respond to this phenomenon, each Member State had to adopt its own set of rules to ensure data protection. However, once companies started collecting data online and reselling it in other countries, it was necessary to establish a regulation that would apply throughout the EU.

The reform had been planned for years, due to pressure from various sectors regarding the uncontrolled use of personal data by governments and companies, and was adopted by both the European Parliament and European Council in April 2016. The GDPR then came into force on 25 May 2018, giving European countries time to make the necessary changes to adapt to the new rules. At the core of GDPR are personal data, both simple and sensitive information about subjects such the ones mentioned before. Article 5, among the 99 contained in the Regulation, sets out the fundamental principles underlying it, to give a general framework of the values in which it operates. These principles are: «lawfulness, fairness and transparency; purpose limitation; data minimisation; accuracy; storage limitation; integrity and confidentiality (security); and accountability» (GDPR 2018). In reality, only one of these principles – accountability – is new to data protection rules.

Finally, and much more importantly for our purpose, the GDPR sets limits to the use of algorithmic decision-making processes. Specifically, Article 22 (1) was introduced to act as an exclusive protection against ADMs, and recites as follows:

> «the person concerned has the right not to be subject to a decision based *solely* on automated processing, including profiling, which produces *legal* effects affecting him or her in a similar way significantly.»

Nevertheless, exceptions to this general wording are made immediately afterwards. One can be subject to an algorithmic decision in three different scenarios. First, if the explicit consent of the person concerned exists, the decision can be made. Secondly, it is allowed in the course of the conclusion of a contract (or its execution), provided that the request submitted by the data subject (i.e. the person whose personal data are collected, stored or processed), has been fulfilled or that there are appropriate measures «to safeguard his/her legitimate interests (e.g. the data subject could express his/her point of view)» (Noto La Diega, 2018, § 45). For example, some law firms use «AI-enabled computer programs to assess the merits of personal injury cases and then decide whether to accept the case or draw up contingency fee agreements. Subsequently, and more generally, algorithmic decision-making may be authorized by law if measures exist to safeguard the legitimate interests of the person concerned» (Ibid.). Typical examples are the prevention of fraud and tax evasion.

Despite its apparent straightforwardness, Article 22 is very complex to interpret. The specific choice of terms used, indeed, opens up to many discussions on what the boundaries of both its applicability and also of the exceptions it lists. According to Noto La Diega, for instance,

> «it is open to debate what *solely* automated means. In the past, it was relatively easy to understand what it could have meant. There was a limited number of organizations taking significant algorithmic decisions and the technologies used were quite rudimental; therefore, reviewing the machine-generated data was relatively straightforward and once a human being reviewed the data, the decision was no longer solely automated» (Ibid., § 53)

According to ICO, «Solely means a decision-making process that is totally automated and excludes *any* human influence on the outcome. A process might still be considered solely automated if a human inputs the data to be processed, and then the decision-making is carried out by an automated system. A process won't be considered solely automated if someone weighs up and interprets the result of an automated decision before applying it to the individual.». (ICO, Guide to the GDPR, 2018).

To clarify the area of application of Art. 22, Art. 29 of the Working Party proposed some examples: «If a human decides whether to agree the loan based on a profile produced by purely automated means, then Art. 22 will not apply. In turn, if an algorithm decides whether the loan is agreed and the decision is automatically delivered to the individual, without any meaningful human input, then Art. 22 will apply» (Working Party Guidelines, Art. 29, 2018). The point, according to Noto La Diega, «is that its interpretation represents a substantial grey area. For instance, it is unclear whether the article applies when the algorithmic system takes the decision, but a human being reviews it» (2018, § 54), and so to what extent a human taking part somewhere in the process has to be considered as an *intervention*. According to Dryer and Schulz,

> «Both the relatively narrow scope of application of the prohibition and the broad range of legal exceptions to the prohibition provided in Art. 22 (2) GDPR – first and foremost on basis of consent given by the data subject – result in very limited cases in which an ADM system is *actually* prohibited. Hence, *partly* automated decisions are going to become a normal part of our everyday digital lives» [emphasis added] (2019).

As a matter of fact, all those automated decision-making processes that are solely concerned with providing a basis or suggestions for a choice that will eventually pass through the human being can be used without exception. Summing up, therefore, to fall under the GDPR rule, decision-making processes must be fully automated, and «must have legal consequences» (Ibid.) or similarly influence the person concerned. If these criteria are missing, the provisions on ADMs introduced by Art. 22 are not applicable.

In any case, it is important to underline that – despite the narrow scope to which they refer – the rules introduced by the GDPR have an important relevance with regard to safeguarding individual rights. However, with regard to social and group objectives, «such as non-discrimination and participation, the GDPR has little to offer» (Ibid.). However, for ADM systems that are «exceptionally» eligible under the GDPR, the Regulation contains legal provisions that can partly safeguard the individual interests of users, such as the indication of a right to explanation. In the next section I will precisely take this right into account, which will result of a particular importance for the final purpose of this dissertation. If implemented, indeed, a right to explanation could reveal the unsatisfactory reasons behind an algorithmic choice or decision for the affected subject. This could be a very useful tool to recognize biases and avoid repeating discriminating patterns in the future.

Lastly, as the only European legal provision currently protecting subjects from the «dictatorship of the algorithm» (Fioriglio, 2015), and considering how long it took to implement it in the best possible way, it is necessary to remember that every word used in the GDPR was weighed and chosen with great care and wisdom.

1.4 A Right to Explanation?

In the period between April 2016 and May 2018, when the GDPR was finally set into effect, Article 22 was analyzed in depth and some clarifications were proposed to the text. Since then, there has been much discussion about the sufficiency of GDPR to tackle ADM systems and whether it was an adequate tool to address AI. Especially, it was questioned to what extent it should have been considered as a powerful strategy to combat the problems arising from the use of technology in society, in close contact with the personal lives of individuals. GDPR was not introduced to specifically tackle misuse of artificial intelligence and machine learning. According to Lilian Mitrou,

> «GDPR does not specifically address AI. Although the difficulties and complexities of digital environments have been taken into account by the designing of the data protection regulatory strategy, the regulatory choice in GDPR consists more in what we perceive as 'technology – independent legislation'. Refraining from technology-specific terminology and provisions seems to be a conscious choice to be attributed to the 'technological neutrality approach'» (Mitrou, 2018)

It is claimed that the GDPR provides the data subject with the necessary tools to properly grasp and challenge an algorithmic decision taken on her behalf. Specifically, it is widely stated (Goodman and Flaxman, 2016)[4] that the subject is guaranteed a «right to explanation», in order to ensure her meaningful information. More generally, it is essential to make clear the nature of this right within the GDPR because it has been considered by several actors a promising and very useful tool in the pursuit of accountability and algorithmic transparency. Some researchers

[4] See also European Parliament Committee on Legal Affairs, 'Report with Recommendations to the Commission on Civil Law Rules on Robotics' (European Parliament 2017) 2015/ 2103(INL) <http://www.europarl.europa.eu/sides/getDoc.do?pubRef¼-// EP//NONSGMLþREPORTþA8-2017-0005þ0þDOCþPDFþV0//EN> accessed 13 May 2020.

have critically addressed it, demonstrating that no right to explanation actually exists within the GDPR. The alleged right to explanation would require data controllers to explain how these mechanisms reach decisions (Wachter et al 2017, 2), which would require them to explain how complex and perhaps inscrutable automatic methods work in practice. Having previously demonstrated the inscrutability of the so-called «black boxes», it would seem a very difficult task to achieve.

In their article, written after the approval of the GDPR, Wachter, Mittelstadt and Floridi critically consider what a right to explanation should include. First, one may refer to two different possible explanations: *system functionality*, i.e. the «logic, significance, envisaged consequences, and general functionality of an automated decision-making system», or *specific decisions*, i.e. «the rationale, reasons, and individual circumstances of a specific automated decision» (Ibid., 3). While the first possibility refers to reporting on the operation and technical processes that are generally expected from the algorithm, the second should provide information about the weight of the characteristics and other circumstances regarding the information processed. Furthermore, another differentiation can be established according to when an explanation is placed in relation to the automated decision-making process. Thus, they identify an *ex ante* explanation that occurs before the process takes place and «can logically address only system functionality, as the rationale of a specific decision cannot be known before the decision is made» (Ibid.), and an *ex post* explanation that occurs after.

The reason why the existence of a right to explanation is inferred has to be found in the combined reading of Article 22 GDPR and comments under Recital 71, Articles 13-14 and Recitals 60-62, Article 15 and Recital 63. Recital 71, in relation to Article 22, requires data controllers to «implement appropriate technical and organizational measures» that «prevents, inter alia, discriminatory effects» on the basis of processing sensitive data (Goodman and Flaxman, 2016). Recitals explain the rationale behind the Articles, but are however not legally binding. Article 22 itself states that:

> «[…] the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.»

As Wachter et al underline, no right of explanation is mentioned here. Data subjects can obtain more safeguards or human intervention, express their views or contest a decision but not to obtain an *ex-post* explanation. What would actually require it, if it was legally binding, would be Recital 71, which states that a person who is subject to an ADM system:

> «[…] should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.» (GDPR 2018)

As it is explained in their research, what is written and specified in recitals provides a guidance to interpret the Article but is not legally binding, and member States do not actually need to transpose recitals in their national law. Unlike other authors such as Goodman and Flaxman, then, Wachter and al do not accept the right that emerges from Recital 71 as a constitutive and fundamental part to interpret the GDPR. In addition to claiming that there is no *ex-post* right to explanation in Article 22, then, they also claim that its omission is intentional. As the authors reported, by looking at the previous drafts and proposals of the GDPR and the negotiations, during which European Commission, Council and Parliament discussed the final text, it clearly emerges how legislators' intentions about safeguards on ADMs and profiling and a legally binding right to explanation were stricter but have been eventually dropped (Wachter et al, 2017, 6). Looking at a report prepared by the European Parliament (EP) in 2013 in response to the original text proposed by the Commission (EC) in 2012, we can see their proposal to extend the application of the GDPR to decisions that were not solely or predominantly automated: they should have actually guaranteed human assessment and an *ex-post* right to explanation by law. EP's suggestion to the draft was to add the word «predominantly» to the Article:

> «Profiling which leads to measures producing legal effects concerning the data subject or does similarly significantly affect the interests, rights or freedoms of the concerned data subject shall not be based solely *or predominantly* on automated processing and shall include human assessment […]»

«With *predominantly* not being adopted in the final text of the GDPR, it would appear the strict reading of *solely* was intended» (Wachter et al., 2017, 17).

Alternatively, it has been suggested that an indication for a right to explanation has to be found in Article 13 and 14 (GDPR 2018), dealing with notification duties and stating that, in the case of profiling, «a data subject has the right to meaningful information about the logic involved». According to Wachter and al, these Articles just require an «*ex-ante* explanation of the *system functionality*», because the notification occurs prior to the decision-making process, «at the point when data is collected for processing». Finally, a similar reasoning applies to Article 15

(GDPR 2018), which sets out a right of access and is considered as the «Magna Carta» of the possibility for data subjects to access information on the collection of their data (Ibid., 7-8). The phrasing of the Article, said the authors, is again future oriented. Two terms especially, «envisaged consequences», suggest that «the data controller must inform the data subject of possible consequences of the automated decision-making before such processing occurs» (Ibid., 9). Furthermore, they argue that, «as with notification duties in Articles 13–14, the GDPR's right of access only grants an explanation of automated decision-making addressing *system functionality*, not the rationale and circumstances of specific decisions» (Ibid.). Finally, «although it is certainly not explicit in the phrasing of Article 22(3), the right to obtain human intervention, express views or contest a decision is meaningless if the data subject cannot understand how the contested decision was taken» (Ibid., 16).

It is certainly debatable whether the language and the choice of certain words are sufficient reasons to question the absolute applicability of the GDPR, and an important role in clarifying these points is certainly played by the courts and the rulings made on the matter. What we can certainly say, and what I am interested in pointing out, is that a strong intention to ban certain technological behavior from the GDPR does not emerge. For this reason, the decision – political, as well as technical – of the use and omission of certain forms in some of its articles, underlines its inadequacy in countering the problems arising from a certain use of technology and AI systems that I will analyze in the next chapter.

Above all, the decision to apply the GDPR solely to completely automated decision-making processes, is highly limiting with regard to the protection of individuals. It means that any human intervention in an AI/ADM system is not regulated within the European Union. As other scholars declared, all a firm needs to do not to be sanctioned or restricted by GDPR, is to introduce any human somewhere in the process, and the firm is no longer basing their decision solely on automated processing. Regarding Member States and the interpretability concerning the vagueness and opacity of these Articles, different national courts have decided to apply different interpretations. The German Court, for example, «has adopted a restrictive interpretation, considering that any minimum human intervention would have excluded the applicability of Article 15 of the Data Protection Directive (which is the former version of Article 22 GDPR)» (Malgieri and Comandé, 2017, 8).

If conceived at this stage in our time, this limitation becomes particularly serious. If we look at all the arbitrary episodes to which individuals are subjected by these systems, the choice of European legislators plays an important role in telling us that they were clearly not yet ready for more courageous choices. Having introduced the functioning of these technologies and framed their European regulation, the consequences of this approach will be explored more deeply in the following chapters, deeper social risks it generates.

In this chapter, I started by researching the particular nature and design of automated decision-making processes, providing some general indications on their technical functioning. Outlining the various stages of the process allows to introduce in the next chapter the different types of biases and especially where they originate. In the second part of this section, the legal framework regulating ADMs in Europe has been presented, with particular attention to Article 22 of the GDPR and its scope of application. Focusing on the alleged «right to explanation», it emerged that proving the actual existence of such a right, which would guarantee the subjects involved an explanation regarding the automated process, is particularly challenging. As demonstrated by Wachter et al (2017), the legal wording in the GDPR refers to an *ex ante* explanation, which therefore cannot reveal the content of the decision in advance. The authors also pointed out that when the Regulation was adopted, there was no intention to include this right. I conclude that we cannot rely on the right to explanation in order to receive a proper account of the algorithmic processes to which we may be subjected. It is therefore possible to proceed with a reasoning on the discriminatory potential of these systems by having a complete idea of the technical context in which they gravitate.

2. Automation and Gender Relations

«In order to design interventions that actually help women, first we need the data.»
– Caroline Criado Perez, Invisible Women

The relation between algorithms and bias has attracted a lot of attention in the last years, leading to a large research that investigates the role of new technologies in perpetuating injustices. There have been plenty cases of discrimination and security problems resulting from system failures to make us reflect on their potential implications. A growing awareness of the effects of biases is emerging, followed by a quest for fairness, transparency and accountability. As I previously explained, automated decision-making processes led by machine learning algorithms can be used in various scenarios. In public sector bodies, they are usually implemented for predictive policing, or for decisions on eligibility for public services like state aid, pensions and unemployment subsidies. In the private sector, on the other hand, some examples can be found in job recruitment, in the granting of loans or other credit services or in the allocation of health care, where privatized.

While the link between gender relations and automated technologies might not appear immediately obvious, it occupies a significant role in the design and implementation of technology. Indeed, «technological design often captures and reproduces controlling and restrictive conceptions of gender, which are then repetitively reinforced» (Collett and Dillon 2019, 4). How technology is deployed and how data is collected and used have a different impact on social groups. Indeed, the two primary ways in which algorithms are biased are race and gender. There have been some cases that particularly attracted the attention of media and research in recent years, also concerning gender discrimination. Among them is the case of Amazon, whose experimental recruiting algorithm excluded female candidates favoring males. This happened because of the historical data used as inputs to the algorithm. It has been proved that to design and educate the software, the appointed engineers used data over a ten-year period, which corresponded prevalently to male résumés. According to Sandra Wachter, they looked «at historical data from the past and at successful candidates, and fed the algorithm with that data which then tried to find patterns or similarities» (Hamilton 2018). But since Amazon employees throughout history were mostly male, the algorithm excluded females from the role of «ideal candidates».

Most of the public cases of algorithmic bias, however, concern systems deployed in the United States and Asia – especially China – and the complete lack of regulation that follows them. Compared to the ones generally adopted within Europe, indeed, other algorithms have much more freedom of action. In these countries, the presumption is precisely that of predicting the future, and to be able to use this key information – indicated by the algorithm that gives a higher risk score – in police departments, courtrooms, and at every step of criminal proceedings. Another example of highly discriminating algorithms are those used for facial recognition techniques, through which huge amounts of biometric data are analyzed and faces are labelled according to different characteristics. These systems have proved to be disproportionately less accurate when identifying women, with results that will be addressed later.

It is believed that with the introduction of the GDPR and the responsibilities it prescribes, Europe is much less at risk of algorithmic bias. Despite the merits of the recent Regulation, however, Europe still shares the same problem with all other countries. The reason for this is that the general mindset on which these technologies revolve is the same, and it has to do with an over-confidence in technological neutrality, avoiding the recognition of the different impact that technology has on different social groups. This is proved by the texts and the choices exposed in the GDPR and will be address in detail. Moreover, as the organization AlgorithmWatch[5] has identified in their 2019 report, cases similar to those reported concerning other countries are actually occurring in Europe as well. DANTE is a project funded by the European Commission within the Horizon 2020 program and uses algorithmic decision making in order to detect terrorism. Automated interviews are carried out with persons that have the intention to cross EU-borders, with an algorithm that needs to determine if they are telling the truth about their case, under the iBorderCtrl system. Predictive policing, then, is apparently being implemented in Belgium and Denmark, where ADMs are also used to assess credit scoring, while in Finland a service is offered to companies that have to deal with many job applicants, assessing worker's personality on the basis of digital footprints[6].

Training algorithms with historical data can be a significant problem when considering public or private services that affect lives. Gender relations, in fact, are constantly changing. They represent a progression that relies on a number of different social factors and historical moments. Using old data, specific portions of reality and history are taken over, that may prove to be wrong because they are subject to evolution. Unfortunately, however, this is not the case with the presence of women in the technological sector, which is still in a clear minority compared to that of men. The fact that the presence of women in IT, Artificial Intelligence and computer science is still – often with worrying numbers – lower than that of men can be identified as a preliminary cause of bias. The strong masculine presence in these sectors, in fact, makes the decision-making process on the implementation and design of technologies unequal, without an adequate presence of voices and experiences representing the interests of society as

[5] AlgorithmWatch is a German based non-profit organization that researches and informs society about algorithmic decision-making processes.

a whole and the effects that this technology can have on every subject. In Europe, based on data collected in 2019, women ICT (information and communication technologies) specialists are only the 17% of the total, and they generally earn 19% less than men. Among the ones working in the sector, «46% of women have reported that they have experienced discrimination in the European tech sector» (WomenTech Network). Then, a study conducted to investigate gender composition by job title for Executive-level positions, found out that «just 1 female Chief Technology Officer out of a sample of 175» (Ibid.). Essentially, most of the technologies that are created today and that are introduced into widely varying and diverse societies are designed by homogeneous groups. It is therefore likely that this pattern representing the gender imbalance in the technological sector partly explains the gender stereotypes perpetrated by ADMs and algorithms. It has also been demonstrated in detail that even in Europe the commitment to the quantitative inclusion of women in these areas (despite being one of the objectives identified for Research & Innovation by the current Commission) has not worked as a corrective to their qualitative exclusion from decision-making (Best et al, 2017).

As Barocas and Selbst underline, «an algorithm is only as good as the data it works with. Data is frequently imperfect in ways that allow these algorithms to inherit the prejudices of prior decision makers» (2016, 671). Most countries in the world have recorded and share a *gender data gap*. This means that most of the data that has been collected in the years were men's, and that there's a significant lack of gender-specific data. The failure to represent a large part of the population leads me to say that data is never neutral. Not on a technological level, nor on a socio-economic or socio-cultural level. Everything from data collection, design of data-driven instruments, and data interpretation fails to acknowledge the gendered dynamics at play. As Judy Wajcman affirmed, «gender relations can be thought of as materialized in technology, and masculinity and femininity in turn acquire their meaning and character through their enrolment and embeddedness in working machines» (2010, 149). What Wajcman refers to seems to recall Judith Butler's philosophy, according to which gender is built through a continuous *performance* in a temporal repetition: «this repetition is at once a re-enactment and re-experiencing of meanings already socially» (Butler, 1990, 191). The nature of algorithms appears to be very similar in this sense: in their case, an epistemology is constructed through the repetition of the same process over and over again. The similarity between the construction of gender and that of technology will be addressed in detail hereafter.

According to the technological functioning explained in the previous chapter, the gender inequality to which I am referring to occurs mainly because of problems regarding the data on which the algorithm is trained, or on which it learns autonomously. But it is fundamental to understand how and why data are biased, which can happen for several reasons.

2.1 The Creation of Biases and Sources of Discrimination

There are various reasons why the output of an algorithmic decision can result as unsatisfactory and discriminatory. For example, discrimination can occur when data inputs about subjects are not relevant enough to reach a correct conclusion, or because the quality – and quantity – of training data do not reflect the complexity of society. Other important factors are the historical context in which the data was generated or the particular forms of measurement error the data contains. In addition to biased data potentially causing problems, the series of choices and practices during the development of the machine learning model, like evaluation methodologies or model design, could lead to unwanted effects. The consequences that arise from these imbalances can be severe: subjects may be deprived of a service or denied access to information.

The word bias is used widely in many different contexts with various meanings. According to Scantamburlo et al:

> «in machine learning, a related concept is used: learning a concept from a finite sample requires making some assumptions about the unknown concept, so as to reduce the search space, and reduce the risk of overfitting the training set. Occam's razor, the principle that the simplest hypothesis is to be preferred, all else being equal, is a classic example of bias in machine learning» (2019, 7)

In this context, it refers to «the inclination or prejudice of a decision carried out by machine learning classifier models which is for or against one person or group, especially in a way considered to be unfair» (Ntoutsi et al, 2019, 4). Acknowledging this problem is complicated, since there is a tendency to believe that the rationality of machines cannot suffer from such errors. When addressing biases, the issue of *fairness* is central. However, there is no one generic approach to it: the analysis on the Art. 22 GDPR has demonstrated how the lack of accuracy in addressing these issues can express discordant interpretations about the same issue. A first distinction to be made is between procedural and outcome fairness. According to Rovatsos et al:

> «Procedural fairness is concerned with the fairness of the steps, input data, and evaluations made in a decision-making process. In a data science context, this could mean an algorithm which processes data about individuals in the same way, regardless of characteristics such as gender and ethnicity. On the other side, outcome fairness addresses the equity of the outcomes of a decision-making process, and how they are distributed across individuals and social groups within the population. It is often discussed in terms of

> discrimination and the denial of opportunities or services to specific groups». (2019)

The fundamental problem is that these different approaches are often incompatible, which means that they require a choice to be made on each occasion to decide on the most appropriate approach for a certain goal. Even whether, for example, an employer is committed to implement procedural fairness to ensure equal treatment for all candidates, it may still find at the end a discriminatory outcome based «on membership of certain social groups or the use of selection criteria (e.g. education)» (Rovatsos et al., 2019, 12). The authors explain that algorithms cannot be optimized towards all metrics of «fairness» simultaneously. Rather, we need an ethical discussion on what reasonably constitutes fairness within specific decision-making contexts (2019, 13).

Different types of biases may be recognized (Götte et al., 2020; Silva and Kenney, 2018), and I will select the main ones to give an overall impression of the algorithmic process and its different steps, each of which may lead to a discriminatory output. *Historical or training biases* are the most researched, being the first step of the process but also being directly exposed to human inputs. When an algorithm is trained on a certain dataset, a foundation for its final decision is created. These sampled data could easily represent historical prejudices and stereotypes, especially since they may have been collected over a long – and diverse – period of time. Furthermore, training data biases can occur when the data initially used to fed the algorithm are poorly diversified, both quantitively and qualitatively. I will go deeper into the nature of this phenomenon, commonly known as data representativeness, at the end of this chapter. A representative case for this type of bias is the recruitment algorithm used by Amazon. When the initial data is already limited according to a certain criteria – in that case the algorithm was trained on a historical data set that excluded women – one cannot hope for a fair outcome. As outlined by Silva and Kenney, revealing such a bias «is nearly impossible in reality because data sources are rarely released to the public» (2018, 15). The only way to find out if a dataset has contributed to discrimination is often in the course of a litigation (Ibid.).

Secondly, an *algorithmic focus bias* can occur. It may arise due to the selection and measurement features and model labels that are often proxies for the desired quantities. These proxies may lead to ignoring important factors or introducing group- or input-dependent noise which results in differential performance (Götte et al., 2020). Deciding to include or exclude certain characteristics from the dataset can have major consequences that developers should absolutely pay attention to. For instance, excluding information such as gender in training an algorithm for health diagnosis can lead to incomplete and adverse responses. Nevertheless, «the inclusion of gender in other situations, like sentencing, can lead to discrimination against protected groups» (Ntoutsi et al, 2020, 4). Yet, some scholars such as Barocas et al (2017) argue that including these variables is the only solution to reach a «fair» outcome. Targeted online advertising is particularly susceptible to this type of bias. Ammit Datta et al discovered that «the use of Google's Ad Settings feature can lead to 'seemingly discriminatory ads'» (2015). For example, they noted that changing the gender setting to female when visiting web pages «resulted in getting fewer instances of an ad related to high paying jobs than setting it to male» (Ibid.). Some of these results, however, are also intentional. «Different content, information, prices, etc. are offered to groups or classes of people within a population according to a particular attribute», such as the ability to pay (Mittelstadt et al. 9). Very often, women are included in this last group of people due to the historical prejudice that characterizes them about working much less than men. According to a recent study conducted by economist Dr. Catherine Tucker and marketing professor Dr. Anja Lambrecht (2018) in 191 countries across the world, women see fewer ads related to Science, Technology, Engineering, and Math (STEM) careers than men. Even whether the targeting of the ad used in their field test was explicitly gender-neutral, it was still shown to over 20% more men than women. As I have already outlined, «such outcomes occur either because those who program the algorithm intend to discriminate or have unconscious biases, or because the algorithm itself will learn to be biased on the basis of the behavioral data that feeds it» (O'Neil, 2016).

Thirdly, *processing biases* are probably the most complicated form to reveal, because coinciding with the «black box». Generally, «developers are not allowed to disclose their source code to the public» (Silva and Kenney 2018, 20). These disfunctions can arise as a result of certain design choices, driven by reasons of efficiency or functionality. Usually, such bias «is created when variables are weighted. Another occurs when the algorithms do not take into account the differences in the cases, resulting in incorrect or inaccurate results» (Ibid.).

Finally, there is a set of *outcome biases* that may manifest once the algorithm has reached a decision. For example, the output could be interpreted according to users' bias. A first problem occurs because too much faith is put in algorithmic decisions, taking them as a fact and not as an indication that needs further interpretation. Another problem with the outcome is that of opacity, closely related to that of black boxes. As I have already pointed out, the reasons for the result could be inexplicable even for the creator of the algorithm or the owner of the software. Another particular example is that of *consumer bias*, illustrated by Silva and Kenney (2018), that reports the example of Tay, a chatbot introduced by Microsoft on Twitter in 2016. Equipped with complex learning algorithms, this human-like bot was given the profile personality of a young American woman. However, Microsoft closed the Twitter account within the same day, after having realized they had created a technological, social, and public relations disaster. Powered by several users, Tay started spreading offensive content, such as «Hitler was right. I hate the jews» and «Humans, Trump will not nuke Europe. I will neutralize him with my terrific wall. Which he will pay for. Believe me. Tay out» (Neff and Nagy, 2016). As Sinders commented after the closing of Tay's account:

> «But if your bot is racist, and can be taught to be racist, that's a design flaw. That's bad design, and that's on you. Making a thing that talks to people, and talks to people only on Twitter, which has a whole history of harassment, especially against women, is a large oversight on Microsoft's part.». (2016, 9)

Among what could manifest after a decision is reached, finally, *design biases* indicate the algorithms that were designed in order to reach a determinate goal but are also used for other scopes. As I previously pointed out, indeed, every machine learning algorithm is built for a precise purpose, and the dataset on which it is trained strongly depends on this purpose. If used in other situations and for different scopes that the initial one, the same algorithm can produce inaccurate and non-optimal results.

What is particularly urgent to stress is that the existence of a bias of any kind is always the consequence of a choice. This choice depends on the «algorithm's design and functionality», that «reflects the values of its designer, if only to the extent that a particular design is preferred as the best or most efficient option» (Mittelstadt et al, 2016, 7). Consequently, «the values of the author, wittingly or not, are frozen into the code, effectively institutionalizing those values» (Macnish, 2012, 158). Often, however, some important aspects are left out. This could also be a consequence of the fact that «software developers are not well versed in issues such as civil rights and fairness» (Silva and Kenney 2018, 12).

2.2 Data Representativeness: «*AI, Ain't I A Woman?*»[7]

> *"He is the Subject, he is the Absolute – she is the Other."*
> – Simone de Beauvoir

One of the most certain and socially worrying causes of general algorithmic bias is that of defects in data representation. Gender – along with all groups that are socially divided into classes or binaries – is particularly central in this respect. The problem of data representativeness, i.e. the number of times a data is repeated in order to be considered representative of a given group, coincides in particular with the stage of data collection. In this sense, overrepresentation or underrepresentation of data may occur for a particular group. The same data, then, will constitute the dataset on which the ML algorithm will initially be trained on, and from which it will not be possible to go back anymore. This is why the collection of data is such a fundamental moment in the programming of each technology. The example drawn on the Amazon recruitment case fits well here. As it has already been clarified, a decision making algorithm can only be as good as the data it is trained on. Another area where data representativeness is particularly relevant in order to avoid discriminatory results is that of facial recognition systems (FR). FR is a very complex technology, which implies multiple ML algorithms at each step of the process. It is now widely recognized that these systems have a much higher error rate in recognizing female faces than male faces. The faces of African American women, then, return more false positives than other groups. Amazon is once again involved, being one of the largest manufacturers and sellers of facial recognition systems. According to a study conducted by MIT researcher Joy Buolamwini, Amazon's FR software *Rekognition* «made no mistakes in identifying the gender of white men, but misidentified women with men 19% of the time and dark skinned women with men 31% of the time» (Buolamwini and Gebru 2018). Buolamwini and Gebru also showed that the word embedding space *Word2Vec*[8] encodes societal gender biases. The algorithms that shape natural language transform words into vectors, and similar words should be close to each other in this vector space. In the space of word embedding, for example, 'man' is to 'programmer' as 'woman' is to 'homemaker' (Ibid., 1). When machine learning models are trained on this data, it may happen that a recruiter looking for 'programmers' who in turn gets help from an algorithm will leave the female curriculum behind.[9]

[7] *AI, Ain't I A Woman* is a poem written by researcher Joy Buolamwini in response to algorithmic gender and racial bias and the inability to recognize female faces of color.

[8] Word2Vec is a set of models that form a two-layer neural network, designed to process natural language. Each word is translated into a vector that represents the semantic distribution of the word in the text.

[9] During the Hacking Discrimination hackathon held at Microsoft New England Research & Development Center on 2017, a visualization tool showing societal bias in word embeddings has been created: http://wordbias.umiacs.umd.edu/

The greatest risk regarding the representativeness of data and their link with discrimination is that of underrepresentation. The quantity of data used to train the dataset is fundamental to have an accurate and diverse representation of the population. In order to be fairly representative, a software should be trained on huge amounts of data, and these data should also be selected following specific inclusive criteria. Yet, «some software is trained on as little as several thousand cases» (Silva and Kenney, 2018, 16). In general, the phenomenon of underrepresentation is not considered as intended, but as the consequence of a lack of material. Actually, in the algorithmic era of Big Data, omitting data about women and their experiences is not a lack of means but a choice. This tendency not only concerns men – although we know that they hold most of the top positions in technology and computer science – but anyone who appeals to an alleged gender neutrality. As Caroline Criado Perez argues in her book *Invisible Women,* «these differences go ignored, and we proceed as if the male body and its attendant life experience are gender neutral. This is a form of discrimination against women» (2019, 15). In her work, she shows that algorithmic polarization is nothing more than the last chain of a process that has very deep historical roots. In fact, the author collects endless examples of social biases towards women, from the medical-health area to the design of cars. In the context of a world designed for men – where women are considered as the *Second Sex* (de Beauvoir, 1949) – it is not surprising that the algorithms developed in this world have inherited these biases. One important element that Criado Perez brings out is the relationship between care work – all those «unpaid caring responsibilities» that typically bind women – and Big Data. As she notes reporting one example of women underrepresentation, «designers didn't know or didn't care about the data on women's unpaid caring responsibilities, the software has clearly been designed without reference to them» (Criado Perez 2019, 284). Moreover, she claims that in addition to representing women insufficiently, these datasets are also distorting and oversimplifying them, resulting in associating women's names with family and housework and men with career (Ibid., 339). Again, «A 2016 analysis of a popular publicly available dataset based on Google News found that the top occupation linked to women was 'homemaker' and the top occupation linked to men was 'Maestro'» (Ibid., 340). The examples are endless, but these are enough to show what it means to be part of an underrepresented group in society, even though it makes up half of it. This problem becomes even more worrying if we think that the same technologies are being used today to allocate social services, concerning health, life, home, children and work.

In her book, Caroline Criado Perez showed how «machines aren't just reflecting our biases. Sometimes they are amplifying them – and by a significant amount» (Ibid., 341). But it is recognition here that plays a key role. Recognizing that this problem exists, that these biases are real and harmful, would be the first step to counteract them. The means to solve this problem exist. The first step, of course, would be to train the algorithms with different data, representative of each group in the same way. Then, more inclusive design choices should be made. There are many examples – mostly from independent realities – of algorithms and software created to detect and contrast these biases. Unfortunately, this is not enough. Solutions exist, not only at the designing level, but also in the realm of politics and law.

This chapter addressed the issue of gender discrimination and its strengthening by algorithms. Different types of biases have been identified, according to the stage of the process in which they originate, showing how each stereotype presented by an algorithmic outcome is the direct consequence of a design choice. The idea of technological neutrality and the metaphysical framework in which it is inserted were then challenged, showing that every algorithmic bias originates in a human social prejudice. For automated processes, however, responsibility is more difficult to determine, and concepts such as equity, intentionality and intelligibility are discussed. Several empirical examples have then been offered throughout the chapter, in order to demonstrate the severity of gender biases and to highlight the importance of such a discourse within a technological reflection, with particular reference to the phenomenon of the representativeness of data. The concept of «technological code» presented by Andrew Feenberg allows to expand these reflections in the next and last section of this thesis.

3. Technology as a Political Phenomenon

«The design of technology is thus an ontological decision fraught with political consequences.»
– Andrew Feenberg, 2002

It is undeniable that most of the examples on biases presented by machine learning and decision-making algorithms are much more explored in other countries. The kind of problems that have been presented, however, show that they are also applicable to the European context. The services offered by Amazon, Google and other transnational companies, for instance, are used worldwide, and the legislative distinction from one country to another is not always easy to address. But even if other countries may suffer from the lack of a legal framework more than Europe, the problem I would like to highlight concerns the underlying shared approach to these technological systems. As emerged from the examples that I have provided, a detailed and conscious understanding of all the biases that may arise at the design level of an algorithm and of the importance to diversify data at the moment of their collection seems to be missing everywhere. As Ntoutsi et al underline, «studies show that representation-related biases creep into development processes because the development teams are not aware of the importance of distinguishing between certain categories» (2019, 10). The process raises a number of different issues, including responsibility, accountability, transparency and accessibility. Accordingly, there has been a growing demand for initiatives to require designers and developers to take these issues into account from the very beginning of the design process by defining ethical and fairness values to be integrated into the system. This trend is also defined as «ethics by design» (Dignum et al, 2018). This lack of awareness can be interpreted both as a cause and as a consequence of what is included in legislative texts. As regards Europe, I have outlined the rationale of the GDPR provisions specifically regulating the use of automated decision-making processes. The choice to omit certain clauses in the case of Art. 22 certainly has consequences for the behaviour of designers, but the choice of exclusion itself is a consequence of not considering biases and technological discrimination as a serious and urgent threat.

Regarding the European policies to be adopted following a discriminatory decision, Ntoutsi et al suggested that the anti-discrimination legislation should apply (Art. 20, 21 EU Charter of Fundamental Rights, Art. 4 Directive 2004/113 and other directives). According to the authors, however, these laws require evidence of prima facie discrimination based on specific prohibited criteria in order to be applied (Ibid., 5). It is difficult, however, to think that the generic anti-discrimination laws, introduced at a time when there was probably not yet much evidence of technological discrimination, could consider the lack of data as a prohibited and discriminatory criterion. To think that generic anti-discrimination law can be applied to a decision taken by an automated decision-making process is misguided. This further demonstrates a lack of recognition of the importance of a detailed regulation of these systems and of their impact on society. The kind of discrimination carried out by intelligent systems and automated decisions is indeed of an unprecedented kind, and it should require a new way to assess these problems. Such a way should address algorithmic discrimination in particular, and not only focus *ex post*, but also address an *ex ante* check in order to judge the appropriateness of data included in the procedure.

The «right to explanation» and the uncertainty concerning its applicability certainly prevent it from being considered as a sufficient tool for people at risk of discrimination. By avoiding to include the possibility to discover the cause of the discrimination, through the unveiling of the algorithm core and thus the opening of the «black box», the GDPR makes it difficult to overcome the discrimination itself. Another fundamental aspect mentioned in the first chapter concerns the decision to specify that the explanation concerns the *system functionality*. Even admitting that opening the black box can be very complicated for reasons that go beyond the law itself, the technical functioning of a technology cannot be considered as sufficient information to understand the complete functioning of an algorithm. The intentionality of transporting the issue from a purely technical sphere to a social ground is completely lacking.

3.1 Technological Neutrality and The Technical Code

In order to deeply understand the political and legislative approach adopted towards technology, it is necessary to analyze the idea of technological neutrality, a fundamental concept underlying ethical and philosophical studies of technology. This idea has received numerous critiques that have developed since the 20th century, especially from the Frankfurt School, and that later became a central interest for science and technology studies (STS), or social constructivism. According to technological neutrality, technological artifacts come into the world in a neutral and objective way, but society then reverses their use or modifies them through its schemes. This implies that every technological instrument is considered to be autonomous, self-regulating and completely decontextualized from society.

Through the examples shown in the previous chapter, especially with reference to the relationship between ADMs and women discrimination, it is possible to identify a tendency to consider technology as neutral. Rarely are technology owners and designers aware of their role in building an algorithm and the potential risk of biases. The reason is that they perceive themselves as initiators of a rational and metaphysical tool that does not need their control. In reality, if a machine presents a bias, it is because it has inherited a corresponding bias from humans. The level of infiltration of human prejudices into a machine is responsible for the discrimination perpetrated by the algorithm. Human error can enter the technological cycle and create distortions in any part of the algorithmic process, from data collection to elaboration to the formulation of the problem. In the field of biases, we can hardly imagine that any machine or calculation system possesses particular opinions that may affect the decisions it makes. On the contrary, human decisions are generally considered to be intentional, and discrimination is managed by making people accountable. One of the greatest technological inconsistencies has always been the vulnerability of algorithms or any automated decision to human beliefs. Technology, as a tool and artifact in the hands of human beings, is nothing but an extension of the existing. In fact, its functioning completely reflects the design according to which it has been programmed, leading to both potentially positive and absolutely negative results.

In the classical conception of technology, efficiency serves as the main principle to determine whether a given technological initiative can be considered successful or failed. But what is efficiency measured on? The degree of efficiency is calculated on a quantitative basis according to a rationalistic approach. Critical theory of technology, however, insists on showing that technology does not act only as a means to obtain a certain response, but also in shaping a way of life. According to Marcuse (1964) «the neutrality of technology places it in the service of the dominant social groups». Langdon Winner investigated technological artifacts in the «way in which they can embody specific forms of power and authority» (1980, 121). To explain this phenomena, he explains that we should not be concerned with the technology itself, but with the socio-economic context in which it was conceived and inserted. (Ibid., 122). According to Winner, this is the premise for a theory called the social construction of technology, which needs as a corrective to those «who fail to look behind technical things to notice the social circumstances of their development, deployment, and use» (Ibid., 122). As Friedman and Nissenbaum suggested, the concept of neutrality «appears to suggest that algorithms are designed in value-neutral spaces, with the designer disconnected from a social and moral context and history that inevitably influences her perceptions and decisions» (1996).

To address the process of technological design and the particular choices that are made, philosopher of technology Andrew Feenberg introduced the concept of «technical code»:

> «A technical code is the realization of an interest or ideology in a technically coherent solution to a problem. […] More precisely, then, a technical code is a criterion that selects between alternative feasible technical designs in terms of a social goal and realizes that goal in design. "Feasible" here means technically workable. Goals are "coded" in the sense of ranking items as ethically permitted or forbidden, aesthetically better or worse, or more or less socially desirable. "Socially desirable"

> refers not to some universal criterion but to a widely valued good such as health or profit.» (2010, 67)

Andrew Feenberg contributed to social constructivism building a philosophy of technology on the political foundations that were missing, to demonstrate that technology is always a «political phenomenon». His approach also aimed at «demolishing some of the most sacrosanct edifices of modern global capitalism's pervasive infiltration of science, rationality and innovation» (2010, x). The technical code expresses «the rule under which technologies are realized in a social context with biases reflecting the unequal distribution of social power» (2005, 47). Feenberg speaks of a «code» because the social construction of technology is always encrypted and never clearly expressed. The metaphor of the black box is central here, often used as a justification for not further investigating the social reasons of a bias. Biases are so difficult to detect also because they are considered as unpredictable side effects, inevitable steps in the process to achieve efficiency. This idea prevents those who follow it from intervening in the technological process to anticipate certain consequences and reverse their course. Moreover, the technical code clearly refers to self-interests, and thus to political preferences. The logic of neutrality has, according to Feenberg, much deeper roots than the contemporary development of technology. It derives from the fact that technological and scientific progress – since the first Industrial Revolution – has been assimilated to an economic and industrial perspective of development. This trend has had many consequences, including that of designing and programming technologies that systematically move away from the empirical conditions of existence. The rationalist infiltration of capitalist dynamics into innovation has accustomed us to associate technology and neutrality, in order to limit interventions to modify it in a more inclusive direction. Using the excuse of an economic achievement through rational instruments, the technical code is covered in neutrality.

Feenberg showed how technical codes always represent certain ideologies, paying particular attention to design choices as political and interested choices. In contrast to the idea of neutrality, then, every space in which an algorithm is implemented is actually a political space. The algorithmic design, the development of its models, the tests made on its performance and the selection of the data to include have direct consequences on the final decision that ADMs are required to provide. Designers' political and social preferences are frozen in the code. Mittelstadt et al, in their reflection on the ethics of algorithms, note that there is no step in algorithmic design that is neutral or linear. Each one requires choices that are not objective, but selected over others (2018, 7).

3.2 Neutrality Legacies in the GDPR

As pointed out by Lilian Mitrou, the approach adopted by the GDPR is that of technological neutrality (2018, 26). Recital 15 has been unofficially titled as «Technology Neutrality», as it suggests that «the protection of natural persons should be technologically neutral and should not depend on the techniques used» (GDPR 2018). Even whether it has established that Recitals are not legally binding, it can still give us another example of an address that is evident elsewhere in the text of the GDPR. The reason for this approach is the intention to focus on the effects of a technology, and not on its nature, in order to not impose any hierarchy. Nevertheless, this appears to be more of a problem than a solution. The first consequence is that of an excessive vagueness, which risks leading to non-existent results. The provisions may be so abstract that they are too difficult to apply and interpret. This is precisely the case of Art. 22 and of the alleged «right to explanation». First, the ultimate choice not to use the keyword *predominantly* instead of *solely* for automated decision making cases where the GDPR applies, has made the application of the article very difficult to define without apparent reason. If these systems have been shown to have a very high discriminatory potential, why not make the applicability of these safeguards more extensive?

Second, as Wachter et al (2017) highlighted, the only legally binding formulation for a right to explanation refers to an explanation of the «logic involved» and of the *system functionality*. This explanation would be provided *ex-ante*, before the algorithm reaches a decision, explaining the technological functioning of that process in general and its broad purpose. After having analyzed many cases of algorithmic discrimination and their opacity, it is clear how such an explanation has to be rejected. A satisfactory explanation would need to be *ex-post*, once there actually is a decision to be challenged. But most importantly, an explanation should provide elements to interpret the algorithmic decision, such as third party interests, design choices and data sets used for training. Certainly, a fundamental role must be given to research, to provide tools to open the black boxes. Trade secrets should be sacrificed, in cases of potential discrimination, to provide an explanation to those concerned and society in general. In this thesis, numerous examples of discrimination against women – one of the most unsolved and debated problems of our time – exacerbated by the use of technological processes have been presented. In this context, without an explanation of why the discrimination occurred, we will never be able to reveal biases and overcome it.

Under this defined approach, it is difficult to imagine that the demands of certain groups are sufficiently taken into account. The choice to exclude certain provisions and formulations that would have add some specifications from the final text of the Regulation cannot be said to be a neutral but a political choice. The alleged impartiality of technological neutrality, in fact, completely hides its political nature. As explained in the previous chapter, the development of technology involves constant choices, and every choice that is made depends on the social visions of those who take it. The neutrality in which the GDPR falls is responsible for the failure to fight gender biases, as well as all other biases. As expressed by the notion of the technical code, the choices of technological design are usually drove by economic efficiency and profit. Some of the examples listed in the previous chapter, such as those about advertising, show how social and economic forces already discriminate, but when implemented with algorithms they can show even worse results. While it is true that humans are biased as much as machines, AI models can embed societal biases and deploy them at scale. Without a direct responsibility in the process of social damage, and hiding behind technological neutrality, there is the risk to depersonalize our behavior in society towards other human beings.

By reproducing and getting used to technological neutrality, we suppress the need for a technological world justified by reality. Moving away from the causes present in the experience, the technicality in which the GDPR falls is not open to explanation. An empirically based explanation would be the only solution. As suggested by Winner and theories of social construction, the focus should be on an «empirical programme of relativism». In this view, «interpretations of technology emphasize contingency and choice rather than forces of necessity in the history of technology» (1993, 365-366).

3.4 Towards a Feminist Explanation

Despite the merit of having introduced a debate on the importance of design choices, social constructivism actually failed in investigating the consequences of those choices. In the field of ethics and philosophy of technology, too little attention has been paid to the repercussions on the personal experiences of subjects and social groups. In particular, «what the introduction of new artifacts means for people's sense of self […] and for the broader distribution of power in society» are not explored (Winner 1993, 367-368).

According to Rosalind Gill and Keith Grint, one of the major theoretical debates of the 1990s has been that on the «relationship between feminism and social constructivism» (1995, 1). Feminist studies of technology (FTS) originated with the second wave of feminism in the 1970s and 1980s and consolidated in the 1990s. Initially, they developed in response to science and technology studies (STS) – or social constructivism – to reveal the patriarchal nature of technology and its influence on its social construction. This field of research investigated technology through different disciplines as a form of masculine power over women's freedom and self-determination. Since constructivism failed to fully satisfy the importance of subjectivity and action, it left out the different impacts of technology on different social groups, falling in the same mistake of technological neutrality. According to Gill and Grint, the attention paid by constructivism to the design of technological means prevented it from seeing women (1995, 18). But as Lohan and Faulkner noted, «since technology and gender are both socially constructed and socially pervasive, we can never fully understand one without also understanding the other» (2004, 319).

Similar to what Caroline Criado Perez showed in the form of examples, some feminist theorists such as Cynthia Cockburn (1983, 1985, 1992, 1993) and Judy Wajcman (1991) have focused on affirming that the alienation of women from technology is a product of the historical and cultural construction of technology as male (Grint and Gill, 1995, 8). In this view, technology represents a patriarchal mode of control and is deeply gendered. The limit of some of the early feminist theories on technology, however, is precisely that of considering technology as a purely and entirely patriarchal instrument. In doing so, they missed the opportunity to make a situated and particular analysis of each technology, to reveal its specific nature and the reasons for its existence. As Gill and Grint well explain: «The issue of how the ideology of masculinity serves to perpetuate women's alienation from, and oppression by, technology remains largely untheorized» (1995, 14). A new feminist approach to technology should retain the positive teachings of social constructivism and feminist studies of the 1990s, but focus on the importance of an «embodied, situated epistemology». As Donna Haraway claimed, «[…] partiality and not universality is the condition of being heard to make rational knowledge claims. […] We do not seek partiality for its own sake, but for the sake of the connections and unexpected openings situated knowledges make possible. The only way to find a larger vision is to be somewhere in particular.» (Haraway 1991, 195). In other words, every algorithmic outcome must be addressed in relation to the empirical effects it produces, rejecting a «once for all and for everyone» interpretation. Addressing technologies in their partiality means that we do not make general considerations about their impact, but that we consider each situation as dependent on the subjects involved, contrasting neutrality. Assuming that technology is not inherently patriarchal, then, we can recognize it in all its situated forms and find new and inclusive ways to integrate it into society.

Despite the weaknesses that have been presented, social constructivism has proposed to attribute real and experiential connections to something that for a long time has been presented as unconditionally neutral and placed outside of experience. In his book *Between Reason and Experience* (2010), Andrew Feenberg sought to bring reason, in terms of technological rationality and its domination, back into a dimension of interdependence with experience. It is essential to reconcile these two aspects once and for all. Any ethics of technology that wants to provide the technological present with a political and legislative direction cannot disregard empiricism. The nature of women's alienation from and by technology has changed profoundly with the advent of automated technologies, and in the case of ADMs, these are systems that can determine access to primary services for women's lives such as health, work or subsidies. Since «algorithmic activities, like profiling, reontologize the world» (Mittelstadt et al 2016, 5), a new theorization is necessary, that rightfully addresses the constantly changing character of technology and that goes beyond the analysis of its patriarchal nature, to address its legislative and political orientation.

A new ethics must integrate new solutions to the static denunciation of patriarchy, in order to improve the technological society. Despite the evidences of male hegemony in the design of technologies, machine learning algorithms are more and more fluid in their immateriality and in their continuous evolution. These are not fixed entities, but instruments that have the power to continuously regenerate themselves with data derived from the social context and are open to progress by definition. As Alison Adam suggested:

> «A view which takes the trajectory of technology for granted also takes the structures of society for granted and leads to the assumption that equality will be achieved without looking to the deeper reasons why the structures of inequality are as they are, in the first place. Too many campaigns to persuade women to enter technical subjects have failed because of their basis in an uncontested liberalism, which fails to scratch the surface of the reasons for inequality.» (Adam 2005, 11)

There are already many creative examples of using machine learning for anti-discriminatory purposes, and hopefully we will encounter more of them in the future. The role for a feminist ethics and politics of technology, anyway, is that of combining a new theoretical thinking with situated practices of intervention. One of the biggest challenges is to push for a new political and legislative agenda, since any ethical reflection cannot ignore these fields of action. To return to the starting point and connect my argument to the first part of this thesis, it is useful to point out that the current legislative framework is moving in an opposite direction to the one I am advocating for. Sharing a neutral perspective when it comes to technological application, the GDPR has not shown enough interest in highlighting algorithmic complexities and putting them at the service of citizens. The opacity, secrecy, inscrutability and complexity of ADMs are kept hidden by an alleged right to explanation. An explanation of the technical functioning is useless to overcome and solve the problems between gender and technology, and increasingly alienates subjects from a true understanding. Consequently, with the maintenance of a neutral perspective, the tensions between gender and technology are not overcome. In order to introduce a full explanation, there must first be a recognition of the problem by legislators and society at large, including the subjects involved, who are certainly not placed in the position of being understood. This will be the task of a new feminist ethics of technology.

CONCLUSION

This thesis was dedicated to exploring the contemporary relationship between the technologies implemented for automated – or algorithmic – decision-making processes (ADMs) and gender bias. In particular, the starting claim asserting the discriminatory potential of this type of technologies has been demonstrated by analyzing different areas of investigation. ADMs, it has been shown, are systems that feed on data that are collected from different social domains and then trained to produce a satisfactory and classifiable response for a given goal. As many of these systems operate without the need of human intervention, since they learn from their past experience, issues relating to accountability, transparency and reliability have been discussed. After a closer examination of the technical functioning of these technologies, in order to better understand where and how the troubles may occur, the relevant European normative framework was presented. This was done in order to provide a complete overview on these systems in society and to evaluate whether the political and legislative approach carried out by the EU is sufficient to tackle potential bias concerns. Dealing with personal data belonging to various social actors, these technologies are regulated under the General Data Protection Regulation (GDPR). The first chapter ultimately intended to prove that the section of the Regulation dedicated to ADMs contains several interpretative problems and vagueness, which do not allow for a clear line of action to be drawn. Most importantly, the legislative framework does not provide the necessary safeguards to understand the occurrence of biases and how to overcome it.

In the second chapter I proceeded to present the specific connection between ADMs and gender relations, seeking to deeper investigate the problem of bias, by differentiating between multiple typologies to detect their different sources and causes. By isolating each step of the algorithmic decision-making process, it is possible to identify from which element of reality the bias has originated. This should serve to have a complete idea of how to intervene against stereotypes. What emerged is that technological design is an extremely important factor to consider when investigating the social causes of discrimination. In this section of the thesis, many empirical examples have been provided to provide evidence of the actual impact of biases, including the exclusion of women from technological jobs, the inability to see certain online advertisements, facial recognition systems that do not recognize female faces, and word embedding algorithms that consider all women as «housewives». For the same purpose, the problem of data representativeness was also analyzed, regarded as the main prejudice-related concern: not enough data are collected on women, compared to those collected on men.

Drawing on the concept of «technological code» proposed by the philosopher of technology Andrew Feenberg, it has been claimed that every algorithmic bias is socially motivated.

From this definition, the third and final chapter has moved into the ethical and philosophical area of investigation, introducing the theories proposed by social constructivists such as Feenberg. Constructivism has been concerned with contrasting the classical idea of the neutrality of technology, a notion that often recurs within this thesis, and its alleged deterministic nature. Despite the merit of having opposed these theories, however, constructivists fail to recognize the importance of the social consequences of technology its different impacts. This is what feminist studies on technology (FTS) have addressed, claiming that technology is deeply gendered and that it perpetrates a masculine culture. The ultimate aim of this thesis was to point out the need for a new ethics of technology, to be feminist in the sense that it addresses gender and other kinds of bias and recognizes the profound need to overcome technological neutrality, and to be deeply situated in our technological present. The challenges posed by artificial intelligence systems such as automated decision-making processes must not be limited to a patriarchal critique of technology and outdated theorizations, but to an extensive study that suggests a new normative and theoretical direction. The emphasis on empiricism is central, especially in demanding new, exhaustive, and inclusive «explanations».

BIBLIOGRAPHY

Adam, A. 2005. Gender, Ethics and Information Technology. Basingstoke: Palgrave Macmillan.

Algorithm Watch. 2019. Automating Society - Taking Stock of Automated Decision-Making in the EU (2019). Available at: https://algorithmwatch.org/wp-content/uploads/2019/02/Automating_Society_Report_2019.pdf. [Accessed on 15 May 2020].

Barocas, S. & Selbst, A. D. 2016. Big Data's Disparate Impact, 104 CALIF. L. REV. 671

Beauvoir, S. de. 1989, c1952. The Second Sex. New York: Vintage Books.

Bowker, G. C. & Star, S. L. 1999. Sorting Things Out: Classification and Its Consequences. Parts II and III. MIT Press: Cambridge, MA.

Buolamwini, J., Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency.

Burrell, J. 2016. How The Machine 'Thinks': Understanding Opacity In Machine Learning Algorithms. Big Data & Society.

Butler, J. 1990. Gender Trouble: Feminism and the Subversion of Identity. New York: Routledge.

Collett, C. & Dillon, S. 2019. AI and Gender: Four Proposals for Future Research. Cambridge: The Leverhulme Centre for the Future of Intelligence.

Criado-Perez, C. 2019. Invisible women. New York: Abrams Press.

Data Protection Working Party. Article 29 (n 9) 8. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2006/wp118_en.pdf. [Accessed 15 May 2020].

Datta, A., Tschantz, M. C., & Datta, A. 2015. Automated Experiments On Ad Privacy Settings. Proceedings On Privacy Enhancing Technologies, 2015 (1), 92-112.

Dignum, V. 2018. Ethics In Artificial Intelligence: Introduction To The Special Issue. Ethics Inf Technol 20, 1–3 (2018).

Domingos, P. 2017. The Master Algorithm. UK: Penguin Random House.

Dreyer, S. & Schulz, W. 2019. The GDPR And Algorithmic Decision-Making. Safeguarding Individual Rights, But Forgetting Society. Völkerrechtsblog, Available at: https://voelkerrechtsblog.org/the-gdpr-and-algorithmic-decision-making/. [Accessed 12 May 2020].

European Commission, Regulation of the European Parliament and the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation) (European Commission 2012) 2012/0011 (COD) Available at: http://ec.europa.eu/justice/dataprotection/document/review2012/com_2012_11_en.pdf. [Accessed 15 May 2020].

European Parliament, Committee on Civil Liberties, Justice and Home Affairs. Report on the Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation) - A7-0402/2013. (European Parliament 2013) A7–0402/2013

Available at: https://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A7-2013-0402+0+DOC+XML+V0//EN&language=en. [Accessed 15 May 2020].

Feenberg, A. 2005. Critical Theory of Technology: An Overview. Tailoring Biotechnologies. 1. 47-64.

————— 2010. Between Reason and Experience: Essays in Technology and Modernity. Cambridge: The MIT Press.

Fioriglio, G. 2015. Freedom, Authority and Knowledge on Line: The Dictatorship of the Algorithm Revista Internacional de Pensamiento Politico - Vol. 10 - 2015 - pp. 395-410 - ISSN 1885-589X.

Frank, P. 2015. The Black Box Society: The Secret Algorithms that Control Money and Information. Cambridge, London: Harvard University Press.

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

General Data Protection Regulation (GDPR). 2018. General Data Protection Regulation (GDPR) – Final Text Neatly Arranged. [online] Available at: <https://gdpr-info.eu/> [Accessed 9 May 2020].

Gill, R. & Grint, K. 1995. The Gender-Technology Relation. London; Bristol, PA: Taylor & Francis.

Goodman, B., & Flaxman, S. 2016. EU Regulations On Algorithmic Decision-Making And A "Right To Explanation". ArXiv, abs/1606.08813.

Hamilton, I. A. 2018. Why It's Totally Unsurprising That Amazon's Recruitment AI Was Biased Against Women. Business Insider. Oct 13. Available at: https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10?IR=T. [Accessed 21 June 2020].

Haraway, D. 1991. A Cyborg Manifesto: Science, Technology, and Socialist- Feminism in the Late Twentieth Century in Simians, Cyborgs and Women: The Reinvention of Nature. New York: Routledge, pp.149-181.

Hill, R. K. 2015. What an algorithm is. Philosophy & Technology 29(1): 35–59.

Information Commissioner's Office, UK. What Is Automated Individual Decision-Making And Profiling? Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-is-automated-individual-decision-making-and-profiling/. [Accessed 15 Apr 2020].

Lambrecht, A. & Tucker, C.E. 2018. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads.

Lohan, M., & Faulkner, W. 2004. Masculinities and Technologies: Some Introductory Remarks. Men and Masculinities, 6(4), 319–329.

Malgieri, G. & Comandé, G. 2017. Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation. International Data Privacy Law, vol. 7, Issue 3.

Marcuse, H. 1964. One-Dimensional Man: Studies In The Ideology Of Advanced Industrial Society. Boston: Beacon.

Mitrou, L. 2018. Data Protection, Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR) 'Artificial Intelligence-Proof'? Available at SSRN: https://ssrn.com/abstract=3386914.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. 2016. The Ethics Of Algorithms: Mapping The Debate. Big Data & Society.

Noto La Diega, G. 2018. Against The Dehumanisation Of Decision-Making: Algorithmic Decisions At The Crossroads Of Intellectual Property, Data Protection, And Freedom Of Information. Journal of Intellectual Property, Information Technology and Electronic Commerce Law, 9–29.

Ntoutsi, Eirini & Fafalios, Pavlos & Gadiraju, Ujwal & Iosifidis, Vasileios & Nejdl, Wolfgang & Vidal, Maria‐Esther & Ruggieri, Salvatore & Turini, Franco & Papadopoulos, Symeon & Krasanakis, Emmanouil & Kompatsiaris, Ioannis & Kinder-Kurlanda, Katharina & Wagner, Claudia & Karimi, Fariba & Fernandez, Miriam & Alani, Harith & Berendt, Bettina & Kruegel, Tina & Heinze, Christian & Staab, Steffen. 2020. Bias In Data‐Driven Artificial Intelligence Systems—An Introductory Survey. WIREs Data Mining and Knowledge Discovery. 10.1002/widm.1356.

O'Neil, C. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York, NY: Crown.

Rovatsos, M., Mittelstadt, B., & Koene, A. 2019. Landscape Summary: Bias in Algorithmic Decision-Making. In What is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?. UK Government.

Scantamburlo, T. & Charlesworth, A. & Cristianini, N. 2018. Machine Decisions and Human Consequences in Yeung, K. & Lodge, M. 2019. Algorithmic Regulation. Oxford: Oxford University Press.

Selena, S., & Kenney, M. 2018. Algorithms, Platforms, and Ethnic Bias: An Integrative Essay. Microeconomics: Welfare Economics & Collective Decision-Making eJournal.

Van Otterlo, M. 2013. A Machine Learning View On Profiling. In: Hildebrandt M and de Vries K (eds) Privacy, Due Process and the Computational Turn-Philosophers of Law Meet Philosophers of Technology. Abingdon: Routledge, pp. 41–64.

Wachter, S., Mittelstadt, B., & Floridi, L. 2016. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.

Wajcman, J. 2004. TechnoFeminism. Cambridge: Polity Press.

————— 1991. Feminism Confronts Technology. University Park, Pa: Pennsylvania State University Press.

————— 2009. Feminist Theories Of Technology. Cambridge Journal of Economics, 34(1), 143–152.

Winner, L. 1977. Autonomous Technology: Technics-Out-Of-Control As A Theme In Political Thought. Cambridge, Mass: MIT Press.

————— 1980. Do Artifacts Have Politics?. Daedalus 109, no. 1 (1980): 121-36.

————— 1993. Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology. Science, Technology, & Human Values 18, no. 3 (1993): 362-78.

Women Tech. Women in Technology Statistics: Where are We? Available at: https://www.womentech.net/en-us/women-technology-statistics. [Accessed June 28, 2020].

Zuboff, S. 2018. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. London: Profile Books.

Zuiderveen Borgesius, F. Council of Europe. 2018. Discrimination, Artificial Intelligence, And Algorithmic Decision-Making. Available at: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73. [Accessed on 15 May 2020].