COLLEGE OF EUROPE
BRUGES
ECONOMICS DEPARTMENT

# Should we let the algorithms decide? A critical assessment of Automated Decision-Making in Europe

Supervisor: Professor Andrea Renda

Thesis presented by
**Eugenia Brandimarte**
for the
Degree of Master of Arts in
European
Economic Studies
Option: European Law and
Economic Analysis

Academic Year 2018-2019

## Statutory Declaration

"I hereby declare that the thesis has been written by myself without any external unauthorised help, that it has been neither presented to any institution for evaluation nor previously published in its entirety or in parts. Any parts, words or ideas, of the thesis, however limited, and including tables, graphs, maps etc. which are quoted from or based on other sources have been acknowledged as such without exception. Moreover, I have also taken note and accepted the College rules with regard to plagiarism (Section 4.2 of the College study regulations)."

Eugenia Brandimarte

# Abstract

The pervasiveness of algorithms in our society, where their automated processes are increasingly replacing humans in many fields of decision-making, fuelled a fierce debate within academia, industry and across regulatory domains. In many European countries, algorithms are increasingly used to determine people's creditworthiness, allocate welfare benefits, predict criminal activity, distribute cases to judges and monitor performance at work. However, many authors are extremely worried by their opacity, lack of transparency, reinforcing effects on inequality and *institutionalised biases* and surveillance traits.

The purpose of this work is to contribute to the European debate on automated decision-making (ADM) by identifying the most relevant challenges and opportunities that its private and public application entails.

Given algorithms' inherent lack of understanding of the social context behind the data, in circumstances where ADM may significantly impact individuals' behaviours, legal rights or opportunities, humans are likely needed to add value to automated decisions, interpreting, explaining and redressing them if necessary.

The European ethics-first approach to AI seems consistent with this human-centric view and, despite the emerging narrative of an 'AI arms race' where speed and profits are favoured over safety and social sustainability, further suggests that Europe wants to run a different race. "Trustworthy AI" would not only be coherent with Europe's founding values but may also prove to be a relevant opportunity for the bloc to set a global standard and thus improve its competitiveness.

When looking closely at three case studies of ADM application in Europe, we notice that, despite all coming from countries with a markedly developed AI strategy, they still suffer from significant weaknesses. First, they reveal that both private and public sectors are not well equipped, to-date, to engage in sophisticated algorithmic processing while still ensuring legitimacy, adequate security measures, transparency and proper redress mechanisms; second, they prove the limited scope of GDPR ADM specific provisions; and third, they uncover the fundamental need for oversight authorities to be conferred enough resources and expertise in order to demand transparency from private companies and perform a forward-looking balancing exercise when it comes to Governments.

Ultimately, before welcoming the effectiveness, preciseness and cost-efficiency brought about by ADM, European countries need to improve their preparedness for limiting and managing its potential pitfalls, both from a regulatory and from a governance perspective.

# Keywords

Automated Decision-Making

Algorithmic fairness

Artificial Intelligence

Ethics Guidelines for Trustworthy AI

General Data Protection Regulation

# List of abbreviations

**AI** – Artificial Intelligence

**ML** – Machine Learning

**DL** – Deep Learning

**GDPR** – General Data Protection Regulation

**AI HLEG** – High-Level Expert Group on Artificial Intelligence

**EC** – European Commission

**WP29** - Article 29 Data Protection Working Party

**EP** – European Parliament

**EESC** – European Economic and the Social Committee

**BEUC** – Bureau Européen des Unions de Consommateurs

**EDRi** – European Digital Rights

**EurAI** – European Association for Artificial Intelligence

**ECAI** – European Artificial Intelligence community

**LAWs** – Lethal Autonomous Weapons systems

**ITIF** – Information Technology and Innovation Foundation

**FCAI** – Finnish Centre for Artificial Intelligence FCAI

**R&D** – Research and Development

# Table of contents

# Introduction

*"People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world."*

(Domingo, 2015)

Since the 1920s, science fiction movies predicted that artificial intelligence (AI) would have broken into our lives in the form of sentient robots, contributing to the prosperity of mankind or, most often, sabotaging or even destroying us. The future we live in now seems instead dominated by a more subtle technological intelligence, which takes the form of algorithms. Algorithms are gradually taking space in our society, where their automated input-output processes are increasingly replacing humans in many fields of decision-making.

An algorithm decides what emails are unimportant to you and thus should end up into your spam folder; an algorithm decides the composition of your Facebook newsfeed and consequently most of the media content you consume online; an algorithm decides the route you take when you type an address on Google Maps; an algorithm decides what jobs are suggested to you as LinkedIn's notifications; an algorithm decides the websites' ranking when you enter a keyword in a search engine; an algorithm decides what to recommend you to watch on Netflix in your spare time.

Based on your personal characteristics and past interactions with the exact service or with other services online, the suggestions coming from the algorithm are perfectly targeted to you, and you are most likely to find them highly relevant.

But algorithmic decision-making is not just a matter of suggestions. In many European countries, algorithms are used to determine people's creditworthiness and therefore to decide whether to grant them a loan or not. In China, an algorithm assigns a social credit score to citizens as part of a wider national reputation system in which the higher their score, the better the services they are entitled to as a reward for their honourable behaviour. In the United States, an algorithm assists judges in bail decisions providing an estimation of the likelihood that a defendant will re-offend when out of jail.

The pervasiveness of algorithms in our everyday life fuelled a fierce and compelling debate not only in many fields of academia, from philosophy and anthropology to psychology, computer science and economics, but also within the whole technology industry and across many regulatory domains.

The purpose of this work is to contribute to such an urgent debate by identifying the most relevant challenges and opportunities that the private and public application of automated decision-making (ADM) entails.

Chapter I aims at offering a complete overview on the functioning of learning algorithms, the frequently hidden dangers embedded in their application and the state-of-the-art strategies to avoid potential biases.

The chapter starts with a preliminary definition of the concepts of AI, ADM, machine learning (ML) and deep learning (DL), which serves as a basis for a detailed analysis of learning algorithms' most problematic features. In particular, the "black box effect" is first introduced, together with some basic models of algorithmic transparency and "explainability". Then, after a brief literature review on algorithmic bias and de-biasing techniques, the most dangerous results of biased systems, such as *feedback loops* and *self-perpetration* of societal biases, are presented in detail. Afterwards, the focus moves to the most debated angle of ADM applications, meaning the concept of fairness, intended as a guarantee that the automated decisions are right, valid and devoid of prejudices. In this respect, the dilemmas posed by the very definition of fairness are outlined, both from a procedural (i.e. *fairness in the process* vs. *fairness in the outcome*) and a statistical perspective (i.e. *predictive parity* vs. *calibration*, *equal false positive/negative rates*). Finally, the novel anti-biases techniques of *causality-based* and *counterfactual fairness* are presented, together with the advantages they bring about.

Chapter II is devoted to the presentation of the state of the debate on AI and ADM in Europe and the critical analysis of the European strategy on AI.

First, the results of a survey conducted on university students and workers aged between 20 and 40 on April 2019 are briefly presented. Despite the inquiry's low scale and limitations, the varied answers are taken as an indicative proxy for civil society's perceptions of ADM and AI. Then, the General Data Protection Regulation (GDPR) provisions specific to ADM are analysed, and their scope and adequacy as safeguards is discussed. Subsequently, the European strategy on AI is described through the presentation of the crucial initiatives on the subject, followed by an overview of the most active institutional and civil society groups that are currently campaigning for increased awareness, transparency and privacy safeguards.

Afterwards, the principle-based approach taken by the High-Level Expert Group on Artificial Intelligence (AI HLEG) of the European Commission (EC), defining trustworthiness as the composition of lawfulness, robustness and ethics, is studied in depth.

In particular, the extent to which an ethics-first approach throughout the development, deployment and use of AI, will be "rewarded" by the market, ultimately constituting a successful strategy for Europe, is extensively discussed. Finally, the possibility for Europe to develop an autonomous model without falling for the emerging "AI arms race" narrative is delineated, together with the possible industrial policy implications.

In Chapter III three heavily debated case studies that perfectly illustrate the possible pitfalls deriving from the private and public application of ADM are examined. These are: (i) the Danish "Gladsaxe model" in which predictive analytics of administrative data is employed for the early detection of children at risk of social vulnerability; (ii) the controversial recruitment technology adopted by the Finnish start-up Digital Minds, that generates a candidate's personality profile based on its "online presence"; (iii) the project OpenSchufa, that highlights anomalies and inconsistencies in the credit scoring algorithm of the German credit bureau Schufa.

The cases are taken as a starting point to investigate whether the private sector is currently well-equipped to engage in the application of ADM and how much adequate oversight, from an expertise and resources' perspective, can be expected from the public one.

For the purpose of presenting an in-depth analysis of the European strategy on AI and of the three case studies, the research activity further involved the collection of interviews from Matthias Spielkamp[1], Stefano Quintarelli[2], Pernille Tranberg[3] and Minna Ruckenstein[4], whose contributions are included throughout the work.

In the Conclusions, the main findings concerning the gains and the drawbacks of ADM in Europe are taken as a basis to sketch a sustainable way forward.

[1] Matthias Spielkamp is the founder and executive director of the non-profit organisation AlgorithmWatch.
[2] Stefano Quintarelli is an IT specialist, entrepreneur, blogger and Member of the AI HLEG of the EC.
[3] Pernille Tranberg is co-founder of the Danish think tank DataEthics and independent advisor in data ethics.
[4] Minna Ruckenstein is associate professor at the Consumer Society Research Centre and the Helsinki Center for Digital Humanities at University of Helsinki.

# Chapter 1

**Artificial Intelligence, Automated Decision-Making and Machine Learning**

The American computer and cognitive scientist John McCarthy first coined the term "Artificial Intelligence" in 1956 for promoting the Dartmouth Summer Research Project on Artificial Intelligence, a summer workshop targeted to experts in various disciplines and aimed at shaping the field's evolution. In his words, AI is the *"science and engineering of making intelligent machines, especially intelligent computer programs"* (McCarthy, 2007).

The modern definitions of AI focus on the capability of such intelligent machines to replicate human behaviour, meaning work, act and react as humans. Some examples of techniques and capabilities associated with artificially intelligent machines are optical recognition, natural language processing, learning, planning, problem-solving and robotics.

The main focus of this work revolves around a particular specification of AI, which relates to computers' problem-solving capabilities and particularly ADM.

According to the January 2019 report edited by the non-profit organization AlgorithmWatch in cooperation with the German broadcaster Bertelsmann Stiftung and with the support of the Open Society Foundations

> *"algorithmically controlled, automated decision-making or decision support systems are procedures in which decisions are initially delegated to another person or corporate entity, who then in turn use automatically executed decision-making models to perform an action"* (AlgorithmWatch, 2019, p. 9).

A first important distinction is to be made between simple algorithms and ML algorithms. In particular, ADM processes may be either built on ruled-based algorithms that analyse the input data against some pre-designed requirements and then return a logic outcome, or on learning algorithms, that infer an outcome based on the statistical analysis of huge amounts of data they are fed into (i.e. the so-called "training data"). In other words, while simple algorithms exhibit a straightforward problem-solving process based on calculation rules to be applied to the input data, learning algorithms flexibly draw a mathematical model from the observation of the training data and then employ such model to produce the outcome.

After having been sufficiently trained, the learning algorithm is able to infer the correct outcome also when confronted with brand new data. This ability is known as generalisation and makes the technique extremely useful in fields where it is impossible to provide specific instructions to the algorithm for the performance of a given task (Bishop, 2006).

As recalled by Burkov, the term was coined in 1959 by Arthur Samuel as a marketing gimmick for IBM to attract new customers and potential employees. Indeed, since even a small modification of the training data is able to significantly alter the desired outcome, the word "learning" is not to be taken in a literal sense (Burkov, 2019).

Based on the input data that are fed into the model, there are three main types of ML: supervised, unsupervised and reinforcement learning.

Supervised learning algorithms operate in a dataset in which both input and desired output are present and have been given specific labels. By observing the labels, the algorithm develops a specific function that associate inputs to desired outputs. When the output of a supervised learning algorithm can take a discrete number of values, the algorithm is defined as a classification one, while when the output is one or a set of continuous variables the algorithm is said be a regression one (Alpaydin, 2014). Some examples of supervised learning algorithms are digit and optical recognition, information retrieval and ranking.

Unsupervised learning consists of giving the algorithm a dataset which contains only unlabeled data and leave the model free to find patterns in such data. In this case, the algorithm groups or clusters the data based on the identified similarities (Wang, 2001). Unsupervised learning is commonly used in market segmentation and social network analysis for the clustering of suggested friends. Finally, reinforcement learning consists of leaving the algorithm free to determine the action that maximises a given reward. Differently from the supervise learning, in this case the algorithm is not given any input/desired output pair but only a reward feedback, which is used to implement a trial and error process (Shutton & Barto, 2017). Some examples of reinforcement learning applications are self-driving cars and algorithms used to play games against a human opponent.

A closing refinement of the ML definition relates to the specific category of deep learning (DL), which may be viewed as a subset of ML with augmented (or more human) capabilities. In particular, while, in case of an incorrect prediction, a simple ML algorithm would require some adjustments or additional inputs from its designers, DL systems are able to train themselves to determine whether their predictions are correct or not, to internalise potential errors and carry on with their automatic assessment (Goodfellow, et al., 2015).

**Appetite for transparency and "explainability"**

One commonly problematic feature of ML is the lack of transparency. In many algorithmic applications, we merely observe the input and the outcome data, without exactly knowing how we went from the one to the other.

Despite there being a broad consensus in favour of algorithmic transparency for processes that significantly impact individuals' rights and freedom, the extent to which such transparency should be granted is still unclear.

On the one hand, greater clarity on how ADM processes function would surely make it easier to spot potential anomalies and possibly correct them. In addition, it would increase people's trust in the automated system and their legal certainty, as it would render feasible to challenge the algorithmically driven decision.

On the other hand, too much transparency could make the algorithm easy to play, especially by people with enough technical knowledge, thus dramatically cutting down its accuracy. Moreover, full transparency would surely violate business secret, that currently protects many algorithms behind widespread services, posing a serious threat to intellectual property rights protection.

Frank Pasquale, who compared our society of deep secrecy and obfuscation to a closed "black box" where, although increasingly monitored by firms and governments, we have any or little knowledge of what is done of our data, argues that extreme transparency measures may be equally suboptimal (Pasquale, 2015). In particular, as insufficient disclosure of an algorithmic process should not be acceptable, nor it should be a hacking initiative aimed at opening a computing system, as it would certainly represent a huge privacy violation for the people involved. In this context, Pasquale demands for a measured degree of transparency, defined as "qualified transparency", which consists of *"limiting revelations in order to respect all the interests involved in a given piece of information"* (Pasquale, 2015, p. 142).

One effective and straightforward compromise could be to ensure transparency by means of a competent authority, empowered to oblige companies that employ particularly sensitive ADM processes to disclose their algorithm and training datasets.

This would require the competent authority not only to have enough resources to perform periodic inspections and monitoring, as the learning algorithm may find new patterns and rationales while being used, but also and most importantly, adequate expertise to successfully audit the system. These two components are among the most crucial challenges that governments willing to engage in a successful AI strategy face.

It should also be mentioned that, at least on the users' side, transparency in itself may not be enough to counteract obfuscation. In particular, releasing the source code that lies behind a predictive algorithm online is not going to increase transparency, as most people would not be able to understand what the code means. For this reason, experts and academics put the

emphasis on algorithmic "explainability", meaning the possibility to clearly justify to users the decisions taken, and the predictions made by algorithms.

Algorithmic "explainability" is the focus of an extensive corpus of literature that attempts to translate ML black box models into intelligible systems, building on previous work on how explanations increase trust and reliance on automated models (Teach & Shortliffe, 1981; Herlocker, et al., 2000; Dzindolet, et al., 2003).

In their work from 2016, Ribeiro, et al. underline the critical importance of trust in ML models used for decision-making purposes, both for the practitioners that are supposed to act on the predictions and for end-users that are subjected to them. As trust is only ensured if the output can be clearly explained, they develop a technique, based on textual or visual aids, that provides qualitative understanding of how the dataset components translate into predictions. Then, through simulated tests together with experiments involving human beings, they show the effectiveness of such explanatory technique in contributing to an increase acceptance of ML models (Ribeiro, et al., 2016).

Similarly, Hendricks, et al. develop a description model for deep visual recognition systems. Focusing on the discriminative features of the observed object, their model predicts a coherent label and describes its appropriateness for the recognised object. Rather than simply defining how the output is determined (so-called *introspection explanation*), the model presents a detailed textual explanation on the visual evidence, defined as *justification explanation* (Hendricks, et al., 2016).

As described above, transparency and explainability are crucial features when it comes to algorithms that significantly impact human lives, both in determining the trust and reliance on them and in counteracting the confusion of increasingly complex technologies. An adequate understanding by companies and end-users of how algorithms work, at least in terms of what data they use and for what purpose they were in principle designed, become even more fundamental in the presence of what we call algorithmic bias, as will be analysed below.

## Data as a mirror: Machine Learning bias and societal reflection

As explained above, learning algorithms draw statistical models and leverage patterns based exclusively on the data they are given as a training, meaning that their decision-making rationale crucially depends on the composition of the input dataset. For this reason, if such dataset exhibits some kind of trend (e.g. prevalence of men over women, prevalence of

individuals with a specific background, etc.) the algorithm tends to replicate the trend, materially transforming it into a bias.

Algorithmic bias can be defined as a systematic distortion in the ML's results that derives from incorrect assumptions of the automated system. In particular, biases can result from designers' personal biases, and thus be built-in in the algorithm, or arise while handling an unbalanced dataset. Even assuming that programmers are not translating their human biases into the code, in most cases, training data would reflect biases that are intrinsically embedded in our society, or *institutionalised*, such as biases towards minorities or sheltered groups.

More specifically, learning algorithms are particularly vulnerable to the so-called *feedback loops*, which determine self-perpetration of biases. This can happen for instance when the number of arrests in a given neighbourhood is used as a proxy for the crime rate in that neighbourhood and then serves as input to a predictive policing model. Such a circular structure is proven to make the algorithm self-sustain its predictions, sending the police repeatedly to the same neighbourhoods, that consequently end up being the ones with the highest arrest record (Ensign, et al., 2018).

One recent incident involving Amazon exemplifies how algorithms may lead to this self-perpetration of societal biases. In 2015, the company discovered that its automated recruiting tool was not gender-neutral. Since the algorithm had been trained by observing candidates' CVs submitted over the previous 10 years which came mostly from men, being the technology industry historically male dominated, the system judged males rather than females as more suitable for technical job positions in the company (Dastin, 2018).

Due to its potentially crucial implications, over the last few years, academic research in the field of algorithmic bias flourished enormously, highlighting the potential discriminatory outcomes of many ML applications.

A project by the MIT graduate student Joy Buolamwini investigates bias in automated facial recognition algorithms in relation to gender and skin type. The project, which aimed at uncovering algorithmic bias in computer visual analysis, what Buolamwini calls the "Coded Gaze", became a research paper, written in cooperation with Timnit Gebru, in which the performances of three commercial systems for image classification are tested and compared (Buolamwini, 2018). For the purpose of the research, the authors create a balanced dataset comprising 1270 images of mixed gender and skin types and then test how the image classification technologies offered by Microsoft, Face++ and IBM perform on it. They find out significant disparities in classification accuracy, showing that all the three systems performed better on males than on females and on lighter-skinned than on darker-skinned

subjects. In particular, the companies performed very poorly on darker-skinned females, which classify as the group with the highest error rates (up to 34%) compared to lighter-skinned males, with maximum error of 0.8% (Buolamwini & Gebru, 2018).

The results show how the lack of diversity in the training data, negligence that we should not expect from commercially sold products, can turn into outcomes that reproduce what may be defined as *institutionalised bias.* As argued by Buolamwini, *"we have entered the age of automation overconfident, and yet underprepared".* This requires companies to improve their accountability by ensuring fairness in the process and transparency (Buolamwini, 2018).

In a famous 2016 paper, Bolukbasi et al. identified an algorithmic gender bias in the context of natural language processing and presented a coherent de-biasing technique. The research looks into the ML technique of *word embedding*, that allows to represent words as vectors, to which other words are associated by analogy (e.g. *"x"* is to *"y"* as *"z"* is to *"w"*). The technique is heavily used by translation services and text autocomplete recommendation engines and also employed in the recruiting field for CVs and cover letters scanning.

The results show that word embeddings trained on Google News articles exhibit significant gender stereotypes (Bolukbasi, et al., 2016). The authors used Word2Vec, the engine employed by Google Translate, to train an analogy generator that would translate from the Turkish language, where personal pronouns are neutral, into English, thus being forced to choose a gender for the translated words. The analogy *"man"* is to *"computer programmer"* as *"woman"* is to *"X"* was completed with the word *"homemaker"*, revealing the presence of a gender institutionalised bias. In general, the word embedding engine would position specific professions towards the extreme points of the *she-he* vector, unveiling the implicit gender associations that are embedded in our everyday language. Concerned by the fact that the reflected bias would amplify or at least reiterate such stereotyped associations, Bolukbasi et al. propose a de-biasing technique based on the removal of the analogy between what should be gender-neutral words, such as professions, and gender. In particular, the authors force the projection of gender-neutral words towards the midpoint of the *she-he* vector, at exact same distance between the two extremes. At the same time, they retain the gender component for gender-specific words in order to preserve analogies such as: *"he"* is to *"king"* as *"she"* is to *"queen"*. Finally, the authors show that their revised algorithm is able to significantly reduce gender bias, while still performing good at analogy and clustering tasks. Despite the fact that the identified bias simply reflect a society trait, they argue that

the de-biasing of word embeddings *"can hopefully contribute to reducing gender bias in society"* (Bolukbasi, et al., 2016, p. 8).

One year later, Caliskan, et al. furher demonstrate that the application of ML to human language captures human-like semantic biases. Starting from the famous Implicit Association Test, which quantifies bias based on the speed an individual needs to associate positive or negative words with specific social categories, the authors create a statistical version of the test to be applied to a huge corpus of online text: the word embedding association test. In order to perform the test, they substitute the measured speed in response time with the distance between words' vectors and show that cultural stereotypes related to race or gender are fully captured by the word embedding technology (Caliskan, et al., 2017).

Despite Bolukbasi's de-biasing approach could in principle be effective also for removing racial stereotypes, the identification of potential biases would require to pre-define specific categories. In this respect, Arvind Narayanan, one of the authors of the Caliskan's paper, highlights the importance of Bolukbasi's assumption that gender is a binary category, while with racial stereotypes the mere definition of categories raises significant problems (Bornstein, 2017).

## Algorithmic fairness: what does it mean and how can it be ensured?

The discussion on algorithmic bias ultimately boils down to a debate on the degree of fairness that a data-driven ADM process is able to grant. However, the concept of fairness is not only difficult to operationalize in practice, but also highly problematic to define.

A first distinction can be made between *fairness in the process* and f*airness in the outcome*.

One way to ensure fairness in the process could be hiding the trait causing bias from the dataset (e.g. gender). This way the algorithm would make decisions ignoring the gender classification of the individuals whose data are contained in the data set ("unaware approach"). However, the removal of a specific attribute does not necessarily imply that the algorithm will dis-regard it. Indeed, in most datasets, information such as gender are completely redundant and can be easily inferred by a learning algorithm, that would translate it into a category. Even if the approach would work, removing a trait revealing an individual belonging to a minority or sheltered group, may end up being naturally discriminatory towards that group (Yona, 2017), failing to ensure *fairness in the outcome*,

In May 2016, the non-profit newsroom ProPublica published an investigation on the software COMPAS designed to assist US judges in bail decisions providing an estimation of the likelihood that a defendant will re-offend within two years from release. The

estimation yielded by the software is based on information such as current and pending charges, history of criminal records, employment and residential stability.

After having analysed thousands of defendants' data obtained after a public-record request from the Broward County in Florida, ProPublica journalists argued that the software was discriminatory against African-American people, overestimating the probability that they would be re-arrested while underestimating the one that white people would. In particular, when the software's prediction was incorrect, meaning classifying high risk a defendant that would not be re-arrested (i.e. false positive) or failing to do so for a defendant that would actually end up being re-arrested (i.e. false negative) black defendants were systematically discriminated against (Angwin, et al., 2016).

When confronted with such claims, Northpointe, the company that developed COMPAS algorithm, responded with a counter analysis of the data. In their claim, they argue that the software was equally good at predicting recidivism rates for blacks and whites, meaning that, given a certain risk score, the estimated recidivism rate was similar for the two groups, a concept they defined as "predictive parity". They add that the higher number of black defendants ending up as false positives simply derives from the difference in base rates of recidivism. In other words, since black people are re-arrested more often, it should not come as a surprise that they have higher risk scores (Dieterich, et al., 2016).

Northpointe's argument is evidently weak. The fact that historically re-arrests of black people exceed the white people ones simply reveal that the training dataset is unbalanced, and it is not a justification for classifying black people as higher risk on average.

Moreover, subsequent work (Kleinberg, et al., 2016; Chouldechova, 2017) points out that the definitions of fairness used by Northpointe and ProPublica were incompatible from a statistical perspective.

In particular, according to Kleinberg, et al. fairness can be formalised in three different stastistical conditions: (i) *calibration within groups*, which implies that, for each of the two groups, when the algorithm rates a set of people as having a certain probability of being re-arrested, a coherent portion of that set of people should indeed be re-arrested; (ii) *balance for the positive class*, requiring that the average score assigned to defendants that are indeed re-arrested does not vary across groups, which would be the case if recidivism rate for white people was consistently underestimated; (iii) *balance for the negative class*, requiring that inaccuracy of prediction does not systematically differ across the two groups, which would be the case if the rate of recidivism for African-American defendants was consistently overestimated. The authors show that, when the two groups differ in the base rate measure,

meaning re-arrest, predictive parity is not only incompatible with calibration condition, but also with the two balancing conditions requiring equal false-positive and false-negative error rates. The application of the same rule to groups with different re-arrest rates, necessarily introduces a bias towards the group with the higher one (Kleinberg, et al., 2016).

The most popular techniques to address the fairness problem relate to what is defined as *causality-based fairness*, which requires to explicitly model unfairness within a causal framework, rather than a purely probabilistic one in which the variables are statistically independent, as it was a causal effect generated by a specific attribute that is causing discrimination (Kusner, et al., 2017; Chiappa & Gillam, 2018; Loftus, et al., 2018).

Kusner, et al. introduce the definition of *counterfactual fairness*, which requires that a decision concerning an individual is equal to the one that would have been taken if the individual was part to another demographic group. This definition is mathematically translated into an algorithm that can actually take into consideration social biases and potentially allow to trade-off between fairness and accuracy (Kusner, et al., 2017).

One year after, Chiappa & Gillam propose a slightly different version of the model, in which the protected attribute is not taken as always problematic in itself, but only *along the unfair pathways*. If more women than men apply to a college with very low admission rates – the authors argue – the fact that the rejection reate is higher for women is not necessarily resulting from a gender bias. Their new approach, defined *path-specific counterfactual fairness*, allows the algorithm to correct the unfair effects of the protected attribute only within an actually unfair scenario (Chiappa & Gillam, 2018).

As already pointed out, the fact that ADM is usually based on easily quantifiable proxies, such as the number of arrests, that do not necessarily reflect what they are intended to measure, crime rate in this case, may determine the entrenching of societal biases, regardless the accuracy of the algorithm's prediction.

In addition, when an algorithm is fed into outcome data that result from human decisions, as is common in the context of criminal justice where past court decisions are part of the input dataset, such outcome data suffer from the so-called *selective labels problem*, meaning that they do not represent a random sample of the population, but a consequence of a human selection. In other words, since the algorithm can only observe the outcome label "breached parole" for defendants that were released on bail in the first place, such labelling is said to be selective and complicates the evaluation of the predictive model (Lakkaraju, et al., 2017).

A decision by a human judge, as Lakkaraju, et al. further point out, may be influenced by factors that are not encoded in the dataset, such as for example whether the family of the

defendant is present at the hearing or not. This "unobserved information" increases the noise in the dataset and may be source of wrong algorithmic predictions.

It is fundamental to remember that an algorithm processes the input data exclusively within their outcome dimension, without actually understanding the meaning behind them.

The lack of meaningfully understanding of the social context in which the algorithm operates seems central in the whole debate on algorithmic fairness. To answer to a famous quote by the mathematician and philosopher Gian-Carlo Rota – *"I wonder whether or when AI will ever crash the barrier of meaning"* – today, it still does not seem the case (Rota, 1985, p. 99).

# Chapter 2

According to a survey[5] conducted on April 2019 on university students and workers aged between 20 and 40 for the purpose of presenting a proxy for civil society's perceptions on ADM and AI, people are puzzled when it comes to trust algorithms. The results, despite significant limitations, (i.e. they come from a small sample and, more importantly, from the same *"echo-chamber"*), are characterized by surprisingly varied answers. Notably, 28.7% of the respondents would trust an algorithm evaluating their CV, while 33.9% would not. 36% would trust one investing their money and 28.1% would not. Answers are more extreme for algorithms giving a medical opinion, that only 16.5% would trust (with above 60% that would not) and algorithmically-driven public transportation, that more than 50% would trust. One interesting feature is the amount of people that answered in the uncertainty. When confronted with the sentence *"I believe AI will impact positively our society"* 37.4% answered *"Neither agree nor disagree"* and when the term society is substituted with "economy", still 33% cannot say. Finally, although almost 80% of the respondents affirm to value the extent to which their privacy is granted and safeguarded when purchasing digital, only 20% of them declare to read the Terms and Conditions of digital products and services. The purpose of this chapter is to outline the state of the debate on AI and ADM in Europe, presenting the regulatory measures currently in force, the most active bodies and organisations framing the discussion and the main steps in the definition of the European strategy on AI. In particular, the approach proposed in the Ethics Guidelines for Trustworthy AI produced by the AI HLEG of the EC, is critically analysed, together with possible regulatory strategies.

### Automated Decision-Making in the General Data Protection Regulation[6]

Since its entering into force, there is a fierce debate on whether the GDPR offers sufficient safeguards from the increasingly elaborate data processing and profiling techniques permitted by algorithms. Article 4 of the GDPR defines profiling as any type of automated personal data processing aimed at evaluating individuals personal characteristics *"to analyse*

[5] The survey was sent via email to alumni from the College of Europe and Luiss Guido Carli Univesity in the form of an online questionnaire and registered 115 answers in total. Despite the low scale and obvious limitations of the inquiry, the extremely varied results are taken as an indicative proxy for civil society's perceptions on ADM and AI. For further information on survey's structure and questions see: https://docs.google.com/forms/d/e/1FAIpQLSfYapCKNkaMDucS_IhJy1o5T69mEuXcY5FZt4K9hei8sBThdg/viewform.

[6] Regulation (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 Directive 95/46/EC (General Data Protection Regulation) OJ L 119, 4.5.2016, p. 1–88.

*or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements"* (GDPR, 2016, p. 33).

Although the definition may seem comprehensive at first glance, many automated processes may not fall under its scope. For instance, applications such as predictive policing, used to determine whether an area needs to be intensively patrolled, may not even make use of personal data as defined by the GDPR[7], nevertheless, as describe in Chapter 1, the use of similar systems may significantly impact the individuals involved, even leading to perpetration of *institutionalized biases*.

In Section 4, Article 22 expressly establishes the individuals' right *"not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."* (GDPR, 2016, pp. 46, §1). The provision, which can be either interpreted as general prohibition or as right to object, revolves around three main points. The first relates to the term *solely*, which narrows the scope of the prohibition to processes that completely exclude human intervention, intended as the involvement of someone that is able to act effectively on the decision, rather than a merely symbolic action (WP29, 2018). The second point concerns the ability of the decision to produce *"legal effects"*, meaning to affect someone's legal rights or legal status, such as benefits' entitlement or denial, conclusion of contracts, granting of citizenship, etc. In this view, an algorithm that automatically generates a decision on whether to grant a loan or not would fit this definition, while a recommendation engine suggesting what to watch on Netflix would not. Last point refers to what *significantly affects* the individuals, specified by the Article 29 Data Protection Working Party (WP29) as something impacting their choices or behaviour in a prolonged or permanent way and even causing, in extreme cases, their discrimination or exclusion. The WP29 further clarifies that even data processing with little impact on individuals, *"may in fact have a significant effect on certain groups of society, such as minority groups or vulnerable adults"* (WP29, 2018, p. 22).

Paragraph 2 states some exceptions: the prohibition does not apply if the decision is necessary for the performance of a contract, authorized by law or based on the individual's explicit consent. In any case, if any of the exceptions apply, appropriate safeguards protecting individual rights and freedom should be granted.

[7] Article 4 defines 'personal data' as any information relating to an identified or identifiable natural person, who can be identified, directly or indirectly.

The limited scope of Article 22, even in its most restrictive interpretation, is blatant. For instance, it does not apply to situations in which an algorithmically computed score is only one of the parameters that a human takes into account when making a decision, such as issuing a credit card or grant bail, regardless of the weight assigned to such parameter and the decision's relevance.

As pointed out in Privacy International report on profiling and ADM in GDPR, it is unclear whether the significant nature of the effect depends of the individual subjective perception or whether an objective threshold can be established. In particular, the nuanced subjective interpretation of "significant effects" given by the WP29 risks placing the burden of proof on the impacted individual and to leave out practices that rely on highly intrusive profiling techniques such as targeted advertising. Ultimately, the *"clumsy syntax"* of the article and the inadequate authoritative guidance on its interpretation, makes its scope rather limited and open to debate (Privacy International, 2017, p. 10).

Concerning the exceptions, as will be further shown in Chapter 3, allowing ADM provided that the data subject has given his consent may permit situations in which there is a significant imbalance of power between the data controller and the individual, that does not have much choice on whether consenting or not.

Looking at Articles 13-15, on which it is stated that the individual must be informed of the possible application of ADM and must be given *"meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing"* (GDPR, 2016, pp. 40-43), some experts claimed the legal existence and the mandatory nature of a 'right to explanation' concerning all processes involving ADM. Nevertheless, many researchers strongly doubt the existence and feasibility of such right, claiming that what GDPR mandates is the provision of rather limited ex ante information on the processing's general functioning ('right to be informed'), rather than an ex-post deep explanation on why a specific decision has been reached (Wachter, et al., 2017).

Obviously, from a privacy Regulation that aims at remaining relevant and applicable in the medium term, we cannot expect an excessively precise language, as, given the speed of technological advances, over-detailed definitions may soon become outdated. Nevertheless, the lack explicitly stated safeguard measures may significantly hamper the Regulation's effectiveness.

Clearly, the GDPR narrow scope and limitations do not reflect the current spectrum of ADM applications already in place, failing to grant, in most cases, transparency of the automated decisions and individuals' right to challenge them. For establishing an effective guarantee

scheme detailing ADM purpose, limitations and related individuals' rights, specific governance tools would be needed, together with complementary regulatory means, and even original ways to apply those already in place.

**Europe's strategy on Artificial Intelligence and Automated Decision-Making**

In April 2018, 25 Member States (MS) signed the *Declaration of Cooperation on AI* to develop a coordinated approach to the technology while progressing the creation of a Digital Single Market. The signatories commit to foster public and private investment in AI, increase R&D efforts and engage in constructive confrontation concerning ethical and legal requirements for the deployment of a "responsible AI". Concerning ADM processes, human centrality, accountability and awareness are advocated (Signatories of the Declaration of cooperation on Artificial Intelligence, 2018).

After two weeks, the EC published a first Communication *AI for Europe* that revolved around three main points: (i) advancing technological research and industrial capacity and creating an "AI-on-demand platform" (AI4EU) that would act as a one-stop shop for customisable tools and services and public sector information; (ii) modernise education and training through interdisciplinarity for supporting the labour market transition and ensuring a smooth adjustment to AI revolution; (iii) develop, through an European AI Alliance, a guiding legal and ethical framework (European Commission, 2018a).

Further delivering on the AI strategy, in December 2018, the Commission published the *Coordinated Plan on Artificial Intelligence* with the objective of progressing cooperation with MS to promote the development, deployment and use of AI "made in Europe". The plan envisages the increase of investments in AI for all MS, required to have an AI strategy in place by mid-2019, especially through European AI public-private partnerships set up in cooperation with European universities, research centres and companies. The Communication further advocates for advancing the creation of a European data space with seamless data sharing, supporting higher education and learning programmes on AI through European scholarships and developing a framework for ethical and trustworthy AI (European Commission, 2018b).

To achieve the latter goal, the Commission appointed 52 experts coming from academia, industry and civil society to the AI HLEG, designated to draft ethical guidelines for ensuring a trustworthy development, deployment and use of AI systems. In particular, the AI HLEG commits to address relevant AI-related challenges such as transparency, fairness and the future of work. As presented in detail later on, the fundamental-rights based approach and

the centrality of ethical principles constitute the foundations of trustworthy AI. In addition, the AI HLEG is required to reflect in its analysis external views gathered from a multi-stakeholder dialogue enabled by the European AI Alliance's consultation platform, through which anyone interested in AI can interact with the group (European Commission, 2018c). In December 2018, following a call from the European Parliament (EP) and as part of a project aimed at building algorithmic awareness (i.e. algo:aware project), the Commission procured a study devoted to the assessment of the most prominent challenges and opportunities behind the application of ADM. The resulting report groups the concerns related to ADM in six main categories, namely fairness and equity, transparency and scrutiny, accountability, robustness, privacy and liability, presenting coherent policy responses from around the world and proposing consistent actions to be undertaken by industry and civil society (algo:aware, 2018).

Two other documents, dating back to 2017, are particularly relevant in outlining the European strategy on AI, that is spurring innovation and competitiveness while maximising the benefit enjoyed by the society. The first one is the European Parliament's resolution on robotics, covering issues such as robots' liability, human employment, safety, standardisation, and ethics (European Parliament, 2017). The resolution contains a particularly controversial proposal concerning the possibility to introduce a specific legal status for robots (i.e. electronic personality), that has been heavily criticized by a group of 285 experts and stakeholders through an alarmed open letter (EU signatories to the Open Letter on Artificial Intelligence and Robotics, 2017). The second document is the opinion on AI of the European Economic and the Social Committee (EESC), which identifies eleven most relevant policy areas, among which ethics, safety, privacy, transparency, accountability, work, education and regulation, and present possible solutions to the related challenges and recommending a human-in-command approach (EESC, 2017).

Many groups and organisations are actively involved in the debate on AI and automation in Europe. In June 2018 the group of European consumer protection organisations – Bureau Européen des Unions de Consommateurs (BEUC) – published an analysis of the impacts of AI on several consumer markets (BEUC, 2018); five months later, the international non-profit organisation AccessNow, published a report focusing on the safeguard of human rights in the digital age (AccessNow, 2018). Along the same lines, the association European Digital Rights works for the development of adequate protection for civil and human rights while closely investigating copyright law, surveillance and net neutrality (EDRi, 2019). The European Association for Artificial Intelligence (EurAI), representative body for the

European Artificial Intelligence community (ECAI) organises conferences and training programmes on AI and sponsors the research in the field (EurAI, 2019). The non-profit organization euRobotics assists the Commission in the development of an effective strategy in robotics for Europe (euRobotics, 2019).

**The Ethics Guidelines for Trustworthy Artificial Intelligence[8]**

The introduction to the Ethics Guidelines for Trustworthy AI ('the Guidelines') starts with an outline of the promising achievements that AI has the potential to facilitate, from confronting climate change and sustainability, to improving mobility and health monitoring, to reduce gender bias. The prerequisites for AI to successfully achieve these goals is human centrality and trust, which should remain *"the bedrock of societies, communities, economies and sustainable development"* (AI HLEG, 2019, p. 4).

Trustworthiness is defined as the composition of lawfulness, meaning compliance with applicable regulatory measures, socio-technical robustness meaning solid security, safety and reliability of AI systems, and adherence to ethical principles based on fundamental human rights, that are the actual foundations of trustworthy AI.

A series of fundamental rights established in the EU Charter of Fundamental Rights, in international human rights law and in the EU Treaties, such as respect for human dignity, democracy, justice and rule of law, freedom, non-discrimination and civil rights, serves as building block for the identification of four crucial ethical principles, namely: respect for human autonomy, prevention of harm, fairness and explicability.

In particular, AI systems should augment the capabilities of human beings, whose oversight should always be secured, and refrain from causing harm, especially to individuals in a vulnerable condition, as what may happen in situation of asymmetry of power or information. Moreover, AI systems should be deprived of unfair bias, to the extent that this is possible, and their purpose and capability should be transparently communicated. The Guidelines acknowledge that the principles may be in conflict with each other under specific circumstances and that, when the nature of the principle in question allows, a balancing exercise should be performed to identify the relevant trade-offs.

Chapter II, devoted to the realisation of trustworthy AI, presents a non-exhaustive list of seven requirements to be assessed by developers, deployers and end users throughout the AI

[8] This paragraph will draw extensively from the *Ethics Guidelines for Trustworthy AI* produced by the High-Level Expert Group on AI set up by the European Commission.

systems' lifecycle: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing and accountability. These requirements should be evaluated through technical and non-technical methods. The first methods pertain to the system's design, architecture, structural mechanisms for ensuring explainability and quality assurance and appropriate testing, while the second ones relate to regulation and codes of conduct, standardisation and certification, accountability and governance, education and awareness, multi-stakeholder dialogue and inclusion.

In Chapter III, a comprehensive assessment checklist composed of more than one hundred questions concerning the analysed AI system is presented, building on all the previously introduced concepts and components. Such detailed framework will serve as basis for a large-scale piloting exercise, to be launched in the summer 2019, to which stakeholders can voluntarily opt in in order to *"operationalise their commitment"* to Trustworthy AI (AI HLEG, 2019, p. 5). Based on the stakeholders' feedback concerning the usability and exhaustive scope of the checklist, the group will produce a revised version in early 2020.

The Guidelines end with an overview of trustworthy AI's opportunities in fighting climate change, enhancing health and well-being, ensuring quality education and assisting digital transformation, and with some examples of so-called critical concerns that may be raised by AI applications. These are: automatic identification and tracking of individuals, that should only be applied if warranted by law or, when consent is the legal basis, if the latter is meaningful and verified; deployment of covert AI systems not adequately disclosed to humans, that should always know, or be able to verify, whether they are interacting with an AI system; AI-enabled citizen scoring, whose purpose and procedure should always be transparent and not in violation of fundamental rights and whose outcome decisions should always be challenged by affected individuals; development of Lethal Autonomous Weapons systems (LAWs).

By the end of June, the group is entrusted to draft a second deliverable, building on its ethics-first approach. This will contain policy and investment recommendations aimed at supporting European development and competitiveness in AI.

**Is there a market for ethics? Guidelines' analysis and possible ways forward**

After having outlined the Guidelines' founding concepts, structure and implementation proposal, it is worth to investigate whether the market is actually going to pay for the additional value of ethics and trustworthiness of AI systems.

According to Ruckenstein, the term "trustworthy AI" is essentially an oxymoron. On the one hand, being an emanation of its developers, AI is as trustworthy as they are, that ultimately depends on socio-technical conditions and is hard to operationalise as system's feature. On the other, the term erroneously sketches AI applications as autonomous agents. Despite there being a market for data ethics, in Ruckenstein's view, it is still unclear whether technology companies will find Trustworthy AI priority enough to invest time and resources for engaging in the structural changes required to position the notion within the organization and to ensure adequate governance. In addition, given the vague nature of the concept, the way in which it is framed is essential, and can go from a genuine attempt to render AI systems more accountable, to what she calls *"ethics washing"*. Furthermore, in the Guidelines, trustworthiness seems largely framed as a business interest, being the concept's translation into practice left to the companies' self-assessment. The multi-faceted debate on AI cannot be left exclusively to the industry and in any case a tick-the-box approach will hardly result in a successful implementation strategy. At the same time, top-down measures such as hard regulation are likely to leave civil society out, while it must be society itself at identifying the public good to be safeguarded within the debate, whether it is humans' well-being, optimization, or a more efficient market economy. To date however, civil society instances are not clearly articulated, as it wants efficiency and equality at the same without understanding that one needs to be traded-off for the other (Ruckenstein, 2019).

Ruckestein is not alone in being sceptic on whether the ethics-first approach is going to pay in the long run. Daniel Castrol vice President of the US think thank Information Technology and Innovation Foundation (ITIF) and Director of the Center for Data Innovation, defined the European approach not only "naïve", but also potentially deleterious for Europe as a whole, which risks falling permanently behind US and China in the AI race. Being consumers primarily interested in products and services' effectiveness, the focus on ethics is unavoidably destined to slow down Europe's competitiveness (Delker, 2019).

As a forward-looking exercise, we could compare the added value for ethical AI to the added value of high privacy safeguards of digital products and services. Although claiming to truly care about privacy, people do not always want or are able to "punish" the companies with low privacy standards. It is true that some users migrated from WhatsApp to Telegram for its more secure technology, however, due to massive network effects, WhatsApp is still steadily dominant (Bucher, 2018). Moreover, although consumers' confidence in big platforms and social media is wavering and *"techlash"* may seem around the corner, trust

in the technology sector as a whole is still very high (Edelman, 2019) suggesting no massive waves of repercussions on the horizon.

Digital users may find algorithmically-driven advertising creepy, but they really value the fact that what they see is relevant for them – Ruckenstein claims. Thus, their vain criticism is not translating into a changed digital behaviour. They feel that they already lost control over their personal data and, most importantly, they do not want to give up the satisfaction coming from Internet's convenience and usability. Small groups are fighting and there is increased awareness concerning the Internet as a public space, the criticism is not mobilizing (Ruckenstein, 2019).

Whether we should reasonably expect the same passive treatment and material disregard for ethical AI, cannot be predicted with absolute certainty.

In her 2016 book, Pernille Tranberg argues that the market development for data ethics is bound to experience the same transition that the market for eco-friendly products and services went through, going from being simply a legal requirement to becoming an investor demand and then a competitive advantage (Hasselbalch & Tranberg, 2016).

She is rather optimistic with regard to what we can expect from the future market's developments for technology products exhibiting an "ethics-added-value" and compares the structural changes brought about the current technological revolution to the greening movement. The GDPR is only into force for one year now, digital literacy needs to be improved and a lot of companies are still dependent on big platforms such as Facebook, thus, a change of paradigm would take time. However, in her view, we will soon assist to a revolution driven from wealthy and educated people, that will move first, as soon as the market for data ethics and privacy secure tools becomes cheaper and more accessible.

She also argues for the establishment of digital products and services that grant higher safety and privacy standards as defaults within the Government administrations. Although they do not benefit from the same amount of network effects as the Tech Giants, if not even European administrations use European companies' digital services as a default, they will never become better and more user friendly. She sees an obligation for all Europeans, being them users, companies or governments, to pay their share for contributing to the development of "responsible AI", a strategy that would definitely pay in the longer run (Tranberg, 2019).

Even in this rather optimistic scenario, as Ulrike Franke argues, if Europe aims at setting ethical AI as a global standard, it should first impose itself as the global standard for AI, rather than lagging behind US and China (Delker, 2019).

In this respect it worth remembering that the Guidelines are not binding nor directly enforceable, making the whole concept of Trustworthy AI, at least for the time being, a mere *"aspirational goal"*, whose ultimate authority critically depends on whether the European institutions will turn it into a binding regulatory framework. For instance, the EU may decide to develop sector-specific legislation, mandating than all AI applications employed in the healthcare sector are Trustworthy, or establish Trustworthy AI as a requirement for technology companies to participate in public procurements. This latter resolution may serve as ground to exclude non-compliant and most probably non-European players from the European market (Renda, 2019a).

Many authors express concerns about the ongoing narrative of an 'AI arms race', the mere perception of which may push countries to *"rush to deploy unsafe (or untested) AI systems"*, transforming their AI potential into a detriment (Scharre, 2019).

The Guidelines' ethics-first approach seems to suggest that Europe wants to run a different AI race from the one that China and US are running. This interpretation is endorsed by Tranberg, according to which, as Europeans we may not be able to fight China and the US on AI in itself, obsessed of being as fast as possible in order to maximise their global influence and economic growth as they are, but we definitely can beat them on "responsible AI" (Tranberg, 2019).

Given the extremely fast technological advancement and high competitiveness of the sector, for a successful translation of the ethics-first approach into practice, the establishment of some form of enforcement mechanism, whether with hard law or softer instruments seems not just desirable (Quintarelli, 2019), but also rather necessary.

Whether Europe will indeed consider using ethical AI as a vaguely protectionist means to re-launch its competitiveness while sheltering its market from US and Chinese competitors is still unclear. In any case, the Guidelines may be a good opportunity for Europe to show the courage to develop its own model for responsible AI, abandoning both the attempt to catch up with US and Chinese big tech competitors and the emphasis on the need for European champions, which seems largely incompatible with both is competition law and data protection framework (Renda, 2019b).

Ultimately, the new race for "ethical AI" may be a test bench to demonstrate to what extent the EU lives up to its founding values and prefers them to a fast growth pace.

# Chapter 3

The objective of this chapter is to investigate some case studies that appear to be relevant for highlighting the possible challenges that the public and private application of ADM entails. The examples are all mentioned in the January 2019 report "*Automating Society - Taking Stock of Automated Decision-Making in the EU"* by AlgorithmWatch and all come from countries with a markedly developed AI strategy: Denmark, Finland and Germany. Given their significant legal and ethical implications, the three cases were heavily discussed by civil society, industry and academia, often hitting the headlines.

**Denmark: The digitalisation reform and the "Gladsaxe" experiment**

With the 'World-class digital service' reform of October 2018, Denmark further establishes its leading position in the digitalization of the public sector.

The reform includes 22 concrete initiatives aimed at fostering the efficient and simplified provision of public services through an increased digital effort, enabling a better collection, management and usage of citizens data by the Government and improving Danes' trust in the public handling of personal information. Among the main initiatives, a Data Council entrusted to issue ethical recommendations and to advance the public debate on data usage and access rights, is established (Danish Ministry of Finance, 2018).

The plan also encompasses the access by private entities, such as financial institutions and insurance companies, to the datasets compiled by the Government since the 60's, where lots of citizens' data were recorded and indexed with personal identifiers. Transparency plays an important role in the Government's digital ambition: accessing the public administration portal *borger.dk*, all Danes would be able to see an overview of the most relevant data the public possesses on them, knowing exactly what public body accessed them and when.

To ensure a coherent development of the digitalization strategy, the Government entered into a pact with Local Government Denmark and Danish Regions, which requires cooperation between Denmark's central administration, regions and municipalities. The reform is backed by an investment fund of 410 million Danish Crowns for the period 2018-2022, aimed at supporting the spreading of new technologies and their public application.

Although the vast majority of ADM processes employed for the provision of public services are considered not only harmless but also synonym of administrative efficiency, many civil society organisations discuss the risks embedded in their application. After the entering into force of the GDPR through the Data Protection Act on the 23rd of May 2018, some minority

centre-left parties of the Danish Parliament complained about the intrusiveness of the Government, authorised, for monitoring purposes, to combine citizens data gathered from a wide array of sources. The group advocated for further-reaching transparency and complete disclosure to citizens of the collection process' details (Mølsted, 2018)

Establishing ethical principles such as human centrality seems a priority for the various organisations active in the field. On September 2018, the SIRI Commission published a report on AI and ethics emphasizing the role of humans in AI environments, the importance of an *ethics and privacy by design* approach and the need for algorithmic transparency (SIRI Commission, 2018) The think tank DataEthics.eu works to promote digital trust through a sustainable use of personal data for a constructive technological development (DataEthics.eu, 2019). The think tank Justitia aims at advancing the debate on technology with particular reference to fundamental rights and the rule of law (Justitia, 2019).

In January 2018, a case in which ADM was planned to be used for the early detection of children in vulnerable circumstances created sensation within both civil society and academia. The 'Gladsaxe model' was presented as part of the wider so-called *ghetto plan*, aimed at fighting 'parallel societies' in 25 residential areas of Denmark. The plan envisages the introduction of special measures in the areas that qualify as "ghettoes" such as the physical demolition of buildings, the privatization of public housing, the maintenance of a more balanced residents' composition, a strengthened policing effort with higher crime-related penalties and the preventive monitoring of families to identify the ones in vulnerable conditions (Danish Government, 2018). The measures, targeted to areas that are mostly inhabited by low-income people with ethnic-minority background, have been heavily criticized for being potentially discriminatory (Bendixen, 2018).

The three municipalities involved in the preventive risk assessment programme are Gladsaxe, from which the model got its name, Guldborgsund and Ikast-Brande. Under the programme, the municipalities would be allowed to collect and combine information from different public sources and to categorise it according to specific "risk indicators" exhibited by a given social context. Such classification would support the automatic detection of children with special needs before they actually reveal their disadvantaged condition.

More than 200 risk indicators, defined as significantly influential on the well-being of children, would serve as inputs to identify families at risk of social vulnerability. To implement the programme, the municipalities would need to combine health and day care data, data concerning the social sphere and employment data. The system would then assign

a specific score to each family based on information such as attendance of doctor's or dentist's appointments, employment status, mental health, divorce and so on. In particular, a parent with a mental illness would score 3000 points less than the base score, a missed doctor's or dentist's appointment would imply a deduction of 1000 and 3000 points respectively, while being unemployed would mean a loss of 500 points.

The Government's intention to deploy such points-based model it in the whole country prompted a strong reaction from the public, which started referring ironically to the mass surveillance traits of the system (AlgorithmWatch, 2019).

Few days after the disclosure of the Gladsaxe model, the implementation of another Government's evaluation system further contributed to the public concern: it was unveiled that some municipalities were monitoring, with no knowledge of the parents, well-being and development of children at kindergartens through targeted data gathering (Kjær, 2018a).

In December 2018, the Gladsaxe municipality was brought again to the public's attention due to the leakage of almost 20000 citizens' personal data including gender, age, CPR number, welfare benefits, family's special conditions and even membership to the Danish church. The leakage showed that the municipality had gathered way more data than those necessary for the purpose of the programme, violating data protection rules (Gadd, 2018).

The massive public criticism and the municipality's negligence pushed the Government to step back from fully implementing the Gladsaxe model. Despite some members of the Government pushing for the initiative to be corrected but maintained, by December 2018, the Liberal Alliance government party's political rapporteur, Christina Egelund, told to the Danish newspaper Politiken that the municipalities were still not best equipped to handle citizens personal data and, as a consequence, the program would have been downgraded. The director of Gladsaxe Municipality's Children's and Cultural Administration referred to Politiken that, although collected, the data had never been used for the purpose of identifying vulnerable children. For the project to be started again, the municipality would need an authorization from the Minister and from the Danish Parliament's Legal Committee and the Social Affairs Committee. However, it is doubtful that the initiative will be re-addressed by the Government before the Danish General Elections of June 2019 (Kjær, 2018b).

Academics heavily criticized the use of public-service algorithms in Denmark, underlining how liberal democracies' quest for efficiency makes them particularly inclined to rely heavily on algorithmic-driven tools. In this context, the democratic infrastructure of the country may not suffice as a safeguard from the risk of turning into a so-called *"algocracy"*, in the harmless attempt to better serve its citizens (Mchangama & Liu, 2018).

Despite the fundamental differences between Denmark's liberal democracy and China's one-party state, at first glance, the Gladsaxe model may resemble China's Social Credit System started by the Chinese government for creating a trustworthy society[9]. Nonetheless, the two models inherently differ not only in their ultimate goals, but also in their main drivers. First, while the final objective of the Danish model is the social inclusion of children from disadvantaged areas, the Chinese model is essentially used to standardize citizens' social behaviour, giving them access to different categories of products and services as a reward. Second, what seems a question of social control in China, seems more a question of money saving in Denmark. Indeed, the Gladsaxe model represents an attempt to rationalise the significant public spending devoted to welfare assistance (Tranberg, 2019).

When reading that *"scoring should only be used if there is a clear justification, and where measures are proportionate and fair*" (AI HLEG, 2019, p. 34) the exact interpretation to be given to the terms *proportionate* and *fair* is rather unclear. It could be argued that the objective's legal legitimacy and public interest might be proxies for proportionality. However, the concept of public interest, and therefore of proportionality, is rather context-specific, dramatically differing between the Chinese and Danish legal frameworks. Although a similar scoring model would be seen as authoritarian in Europe, it is worth remembering that the Chinese Social Credit System, instead of being contested, seems to be largely appreciated by the citizens, that possibly consider the Government acting in good faith and the building of a more disciplined society in the whole country's public interest (Spielkamp, 2019; Tranberg, 2019).

According to Ruckenstein, despite the social infrastructure of the two countries being so different, the logic behind the scoring algorithms, that is predicting individual problems from the observation of socio-economical characteristics, is equally punitive. A system that would look at collective patterns, such as aggregate health determinants for improving society's health conditions would be dramatically different. In any case – she continues – *"if you are talking about ghettoes you lost already"* (Ruckenstein, 2019).

Another element of critical importance when assessing proportionality may refer to the actors that collect and access the gathered data. Supposing a scoring model is used to measure whether a Government should increase the medical effort in some communities

[9] The Chinese Social Credit System employs big data analytics to assess the economic and social reputation of Chinese citizens and businesses through a score reflecting their trustworthiness. Compliance with the law and with accepted social norms affects the score, base on which the subject is given access to certain rights, such as booking a flight or train ticket.

(e.g. enhancing the provision of psychological help services, setting up monitoring programmes to early predict depression, etc.) the data would be exclusively accessed by healthcare professionals, required by law to adhere to a specific code of conduct. A similar framing would presumably ensure legitimacy of the process as well as its proportionality (Quintarelli, 2019).

Despite the noble intent and cost efficiencies, a profiling system based on the indiscriminate collection on family data essentially amounts to mass surveillance. Moreover, in profiling children who come from poor and socially problematic residential areas we run the serious risk of *"putting them in boxes"* rather than helping them, while entrenching a prejudicial pattern (Tranberg, 2019). Indeed, as shown by Chouldechova, et al., the use of administrative data for performing predictive analytics in the field of children welfare may determine a more frequent inspection of families that are already poor or on welfare, emphasising, as a consequence, an already existing bias (Chouldechova, et al., 2018).

As long as ADM suffers from structural weaknesses such as imbalanced and incomplete datasets, it is wiser to use algorithms as a mere add-on to human-decision making. As argued by Tranberg, the excessive and too immediate reliance on algorithms would not only make the automated systems prone to mistakes, but also imply a rapid loss of jobs, with their crucial role of social equalisers. In order to give ourselves the time to re-educate humans that are to be replaced by algorithms, *"our democracy has to decide the pace, not technology"* (Tranberg, 2019).

The Gladsaxe model perfectly illustrates the fundamental need for societies and administrations where ADM is increasingly pervasive to maintain a proper balance between the quest for efficiency and the safeguard of civil liberties. The opportunities deriving from technology's augmented capabilities, especially if employed to serve the public interest, should not be overlooked. At the same time, adequate resources and expertise in the public sector are essential in order to ensure the performance of a robust and forward-looking balancing exercise.

**Finland: The Finnish AI Programme and the start-up Digital Minds**

At the end of May 2017, the Finnish Minister of Economic Affairs launched an AI Programme to boost economic and digital growth in the private and the public sector. Few months later, the working group appointed for deploying a coherent strategy published a report containing an overview of Finland's stance on AI, together with several

recommendations to ensure Finland's leading position in the field (Finnish Ministry of Economic Affairs, 2017).

The report underlines the positive impact of AI on economic growth and the necessity for Finland to readily adapt to the current technological transformation. A field study shows that, due to its high degree of digitalization, education level and peculiar business structure, in a list of 11 developed countries, Finland figures as second after the US in terms of AI-related growth potential (Purdy & Daugherty, 2018). The approach of the Minister is extremely positive. Private businesses are encouraged to become pioneers in AI, with the promise of significant market rewards, while the Government is given the opportunity to respond in a more efficient way to citizens' needs, even predicting them beforehand.

The only sketched concern is the potential loss of jobs caused by increased automation and the consequent uncertainty in the labour market. However, the report underlines, AI developments will likely spur the demand for high-skilled workers, thus also creating new job opportunities.

The urgency for Finland to invest in technology and AI as main drivers of economic growth relates to the limited growth potential of its other production factors (i.e. labour and capital) and its small internal market. Technology-intensiveness is seen as the key asset for a sustainable growth of the Finnish economy, which is estimated to double by 2030. (Finnish Ministry of Economic Affairs, 2017).

It is interesting to notice that privacy-related concerns are not covered by the report, where the word "privacy" only appears twice and always linked to the strengthening of security measures rather than related to the protection of citizens' private life. According to Ruckenstein, this may derive from Finnish people's strong trust in the Government, from which they do not feel they need protection. For this reason, the privacy debate in Finland is way less politicised than in other MS and mainly focused on US companies' practices (Ruckenstein, 2019).

In order to promote AI literacy, the online course "Elements of Artificial Intelligence" has been created, in which more than 100000 of people enrolled. In addition, the interdisciplinary Finnish Center for Artificial Intelligence was founded with the objective of bridging the gap between technology's technicalities and the impacted people (FCAI, 2019).

Over the last few years, Finland demonstrated to be very sensible to issues of non-discrimination within the application of ADM. In 2015, the National Non-Discrimination and Equality Tribunal started an investigation into the credit scoring methods applied by the company Svea Ekonomi (AlgorithmWatch, 2019). The company had refused to grant a loan

to a 30 years old Finnish-speaking man living in a rural area based exclusively on statistical considerations. In particular, although the company had no data on the man's prior payment history, he was given a low credit score based on factors such as his gender and mother tongue (i.e. failure to repay loans is more frequent for man than women and for Finnish-speaking people than for Swedish-speaking ones). In April 2018, Svea Ekonomi was prohibited to reiterate its practices and imposed a conditional fine.
Access to data is central in Finland. In 2016, along with the introduction of the GDPR, MyData initiative was started with the objective of promoting a new approach to personal data management, combining fundamental rights, human centrality and business needs (MyData, 2019). Shortly after, the open community MyData Alliance was established to spread the development of MyData-centered services among businesses and start-ups.

The Finnish start-up company Digital Minds, started in 2017 by two young Finns, proposes an automated personality assessment technology for recruiting purposes (Digital Minds, 2019). Specifically, the company offers a platform in which the candidate directly enters his personal email's and social network's credentials (i.e. Gmail and Microsoft Office 365, Twitter and Facebook) in order to have his whole online presence analysed. The automated processing of the language used in different online contexts, the reactions to emails and posts and the digital interaction with others, allows the tool to compile a personality profile of the candidate based on some standard personality traits: openness to experience, conscientiousness, extraversion, agreeableness and neuroticism. The assessment is performed through a repackaged version of IBM Watson's Personality Insights engine and sometimes complemented with the software for speech and facial expression analysis HireVue.
According to the company, the tool grants a cheaper, faster and more reliable assessment with respect to traditional personality tests, of whether an individual is "a good fit" for the organization which is considering hiring him. Indeed, it could be argued that a similar technique removes the bias that pushes candidates to answer to personality tests based on what they think the company is expecting, influenced by the so-called "social desirability bias". In addition, during a recruiting procedure, personal considerations of the examiner can easily get in the way of a fully objective and unprejudiced assessment.
The company has been inconsistent in disclosing the number of clients it actually serves and, more importantly, the number of candidates that refused to undergo the automated

processing. In addition, the details concerning how the whole processing works or what safeguard measures are implemented are still unclear (Ruckenstein, 2019).

In any case, candidates are obviously in a weaker position with respect to the employer organization and their consent is arguably not completely "free". Nevertheless, given Article 22's limitations outlined in Chapter II, the GDPR is doubtfully adequate to tackle such a clear power asymmetry between the data subject and the data controller.

First, it is uncertain whether accessing private emails for the purpose of creating a personality profile amounts to collection of personal data. Second, if the hiring decision is not based exclusively on the automated psychometric analysis, is difficult to argue that the processing falls within the scope of Article 22. As a matter of fact, the company claims that the tool does not intrinsically qualify as ADM, since it merely automatises a process based on which the company then makes the decision. AlgorithmWatch is currently waiting for the Finnish Data Protection Ombudsman's opinion on the company's practice, that was supposed to be published by the end of January (AlgorithmWatch, 2019).

Regardless the potential of removing candidates' and recruiters' biases, the practice seems problematic from both a legal and an ethical point of view. Since the detailed functioning of the process is still unclear, some questions may be helpful to determine whether safeguards measures are in place to make it legal and remotely ethical.

A first question relates to whether the candidate is adequately informed on the procedure's functioning and always given the possibility to undergo the standard personality tests, thus refusing to consent to the automated processing. In addition, it is highly relevant whether the analysed data are only transiting into the system or whether they are kept for future predictive purposes. If they are never stored by the company, never visualised by a human and immediately erased after being scanned, the process would amount to an entirely machine-driven collection, to which security requirements can be quite easily imposed.

Certainly, an auditing assessment by an accountable body such as the Data Protection Authority may serve as relevant guarantee of compliance with applicable data protection regulation. Hopefully, the opinion from the Finnish Data Protection Ombudsman will shed some light on the case, revealing whether the practice doubtless raises concerns and to what extent the popularity and level of public attention on the company was inflated by the fact that it was mentioned in AlgorithmWatch report.

In any case, many authors claim that humans are still much needed in the hiring field, whether as grantors of transparency, explainability and compliance with the law, as auditors or as effective decision-makers (Spielkamp & Kaiser-Bril, 2019).

**Germany: The Federal AI strategy and the project OpenSCHUFA**

In November 2018, the Federal Government's AI strategy was launched as part of the wider Germany's digitisation strategy ("Hightech- Strategie 2025). In detailing the strategy, the Government acknowledges AI's great potential and economic significance, together with the urgency to develop a holistic framework for its future advancements. Such framework is intended as open to continuous adjustments, based on instances coming from politics, science, industry and civil society. (German Federal Government, 2018). In a nutshell, the strategy aims at creating a high-quality brand for "AI made in Germany" and revolves around three main goals. The first is to safeguard the country's future competitiveness through making it a leading location for the sector's future expansion. This is planned to be achieved with a strengthened R&D effort at home and an increased international cooperation, through the enhancement of the Franco-German research and development network. The second goal concerns the responsible development of AI, that should always serve the common good, especially when its application significantly affects individuals' lives. In order to grant adequate oversight and to promote the dialogue on human-centric AI, the German observatory for AI is established. The third goal relates to the successful integration of AI in society, from an ethical, legal, cultural and institutional perspective. With an informative objective, the Plattform Lernende Systeme aims at fostering cooperation and mutual exchange among different stakeholders, pooling expertise from science, industry and civil society (Lernende Systeme, 2017). As in Denmark, a Data Ethics Commission empowered to issue ethical standards and guidelines is set up.

Several organisations and business associations have already contributed to the debate on ADM with a number of analyses and position papers. In representation of German digital companies, Bitkom recently published a report on the social challenges deriving from AI and ADM, focusing on accountability, ethics and regulatory implications from an industry perspective (Bitkom, 2019). On behalf of Germany's science professionals, The German Informatics Society produced a detailed analysis of algorithmically-driven scoring systems, together with relevant policy proposals such as algorithms' testing and probe, standardization as a means of explainability, increased awareness and data literacy (German Informatics Society, 2018). The non-profit research and advocacy organisation AlgorithmWatch campaigns for increased intelligibility and auditing of ADM processes, especially if aimed at predicting or influencing human behaviour or yielding decisive outcomes (AlgorithmWatch, 2019).

In Spring 2018, the organisations AlgorithmWatch and Open Knowledge Foundation Germany started the project OpenSCHUFA with the aim of shedding some light on how the private credit bureau SCHUFA computed the credit scorings of about 70 million people in Germany. Based on individuals' comprehensive financial history, including paid and unpaid bills, bank accounts, credit cards, loans, fines, possible legal proceedings and judgements, SCHUFA computes a personal credit scoring ranging between 0 to 10000. Such scoring is adjusted based on different business segments and then used by private companies as a proxy for potential customers' creditworthiness, thus representing a "passport" for accessing a wide range of products and services. It is among the decisional parameters for issuing mobile phone contracts, granting bank loans, delivering rental services and allowing payment on account in online purchases. With a low SCHUFA score the ordinary participation in social life is significantly compromised. Given its huge market share in Germany, around 90% depending on the sector, SCHUFA has an enormous power of affecting individual lives.

The credit bureau has always been highly contested by the press, which would report cases of incorrect or incomplete datasets, mistaken identities, late or absent data updates, bouncing of responsibility between SCHUFA and the banks, etc. The frequent recurrence of such events encouraged Matthias Spielkamp and his colleague Lorenz Matzat to start the project OpenSCHUFA, aimed at checking for systematic problems in the creditworthiness' determination (Spielkamp, 2019). First, the data subjects were requested to contribute to data crowdsourcing through the donation of their credit reports, from which personal information were deleted, and through the compilation of separate questionnaires for the purpose of checking whether some personal characteristics were systematically driving discrimination. About 2000 reports were analysed by a group of data scientists gathered by AlgorithmWatch and Open Knowledge Foundation, together with data journalists from the broadcaster Bayerischer Runkfunk and the news website Spiegel Online.

Although the data are not representative of the whole SCHUFA database, women are significantly less than men and elderly people are underrepresented, the analysis, published in November 2018 by SPIEGEL Data and BR Data, revealed a series of anomalies in the scoring system.

First, it worth mentioning that, for almost 25% of the individuals recorded in SCHUFA database, the credit bureau possesses a maximum of three information entries related to the person's business life, such as the conclusion of a mobile phone contract, the issuance of a credit card or the starting of a bank account. Interestingly, the analysis showed cases in which, even in the presence of positive indicators only, the score eroded in time and the risk

increased with respect to the base one. In other words, SCHUFA rated negatively a number of people without actually having negative information on them (e.g. payment failures, debt defaults, etc.).

The investigation also underlines that changes of bank account provider are often causing a lower score, which is coherent with SCHUFA's long history of information losses.

Another interesting finding relates to the significant discrepancies while applying different versioning of the scoring algorithm. In some cases, the results obtained using "Version 2.0", which includes personal information such as age and gender, and the ones obtained from "Version 3.0", which only takes into account business life information, differ by up to 10%. Despite SCHUFA itself declared on its website that the updated 2017 version is more accurate, a lot of companies still employ the previous versioning. Indeed, throughout 2018, the majority of orders were for "Version 2.0" (SPIEGEL Data & BR Data, 2018).

After OpenSCHUFA results were published, many asked for increased transparency by the agency, whose algorithm for computing credit scorings is considered a trade secret, thus not publicly available. On January 2014 the Federal Court of Justice confirmed this view, ruling that the credit bureau was not required to disclosed how its algorithm exactly worked or weighed different personal characteristics (Hunton Andrews Kurth LLP, 2014). In particular, the court did not consider SCHUFA's exact formula to be necessary for enabling citizens to challenge their score.

The score computed by the US data analytics company FICO is often taken as a good example of a transparent scoring, since it precisely states what goes into the model and how every parameter it is weighted. The economist Gert Wagner, member of the Expert Council for Consumer Affairs (SVRV) and advisor to the Ministry of Consumer Protection, declared that algorithms' transparency is the only way to algorithms' fairness (Albert, 2018).

SCHUFA's argument according to which complete transparency increases the risk of system manipulation is valid. Indeed, transparency on the regular users' side is not necessarily needed, as long as that there is full disclosure to the oversight authorities (Spielkamp, 2019).

It could be claimed that information such as residential area, age and gender should be removed from credit scoring models, as they are not predictive of payment behaviour and can lead to discrimination. However, German privacy law already regulates for what categories are allowed and does permit their use, provided that the scoring is not exclusively based on personal characteristics.

The OpenSCHUFA story uncovered the need of more effective and accessible redress procedures to be made available to individuals that are willing to challenge their SCHUFA

score. These could take the form of alternative procedures to be triggered upon request such as the possibility to submit additional documents to be reviewed by a human being or the testing of a given decision in redundancy system where the individuals in the datasets possess different protected attributes(Quintarelli, 2019).

Whether, as happens in France and Belgium, credit scoring activities should be left to the public sector, given their potentially huge impacts on individuals and the consequent need for an accountable actor, is still unclear. According to Spielkamp, if regulation and oversight mechanisms work properly, private companies should be allowed to do it.

When the OpenSCHUFA journalists submitted their findings to SCHUFA and asked for comments, the company replied with a nine-page letter, but did not allow them to make the answer public. Moreover, it seems that the credit bureau asked an outside group to assess the credit scoring algorithm and submitted a report to the Data Protection Authority, which in turn did not make it public.

In order to tackle algorithmic discrimination efficiently and act aggressively when needed, especially in cases that are largely concealed from the public, the oversight authorities would need enough resources and expertise, which does not seem to be the case for the German Data Protection Authority (Spielkamp, 2019).

# Conclusions

The increasing ubiquity of decision-making algorithms in humans' everyday life perfectly exemplifies another of the dilemmas that technological advancements impose on civil society. While we may not even notice that some systems we interact with are fully automated, many ADM application have the potential to significantly impact individuals' rights and freedom. Despite some authors arguing that humans are even more biased and unreliable as decision-makers than automated systems (Miller, 2018), the vast majority of experts, academics, and journalists are deeply worried by their opacity and lack of transparency, reinforcing effects on inequality, surveillance traits and potential takeover on humans (Pasquale, 2015; Zuboff, 2019; Eubanks, 2018; Kissinger, 2018).

As shown in Chapter I, there is plenty of evidence that, if trained with real-life data, learning algorithms can easily pass-on *institutionalised biases* into their outcome decisions, reinforcing them in a self-perpetrating process. Being societal biases essentially a human evolutionary trait developed in a specific historical and social context, it could be argued that, in order to ensure fair and ethical ADM systems, it is society that should be de-biased, rather than datasets. While this is a fundamental goal to be pursued in itself, the risk that biased algorithms may contribute to the building of a pervasive digital infrastructure where their worst features persist in time, should not be underestimated (Bornstein, 2017).

As already pointed out, the lack of substantial understanding of a given social context confines learning algorithms inside the barriers of meaning (Rota, 1985) and relegates them to mere tool of analysis, albeit a very powerful one.

As a consequence, in circumstances where there is valuable meaning behind the data grid and it does not exclusively serve a mechanistic process, humans are likely needed to add value to algorithms' decisions: interpreting, explaining and redressing them if necessary.

The European ethics-first and human centric approach to AI seems consistent with this view. In particular, as argued in Chapter II, despite the emerging narrative of an 'AI arms race' in which countries favour speed and economic growth over safety and social sustainability, the AI HLEG Guidelines suggest that Europe wants to run a different race. "Trustworthy AI" would not only be coherent with Europe's founding values but may also prove to be an opportunity for the bloc to set a global standard and thus improve its competitiveness in the technology sector.

Whether the market is going to pay for the added value of ethics cannot be predicted in advance. In this respect, the transition experienced by green and eco-friendly products, going

from simple results of legal requirements to proper customer's demands, constitutes a positive indicator on how the concepts of sustainability and responsible growth may enter consumers' utility functions. When looking, as informative example, at the extent to which consumers value privacy when purchasing digital products and services, the picture is slightly less encouraging. Digital users are not always able or willing to stop relying on companies with low privacy standards, especially if their products are highly usable, effective and convenient. However, it could be argued that, in contrast with the "older" greening movement, we are now living just the first stage of the data ethics revolution, that would take time to gain supporters. Provided that digital literacy is continuously improved, and awareness is raised in effective and meaningful ways, we can reasonably expect ethics to become a valuable (and paid for) feature in European digital markets.

When looking closely at three heavily discussed cases of ADM application in Europe, we notice that, despite all coming from countries with a markedly developed AI strategy, they still suffer from significant weaknesses. In particular, both the German and the Danish cases reveal that not the public nor the private sectors are well equipped, to-date, to engage in sophisticated algorithmic processing while still ensuring legitimacy, adequate security measures, transparency and proper redress mechanisms. The Finnish case on the other hand may be taken as a real-life proof of the limited scope of GDPR ADM specific provisions and of the framework's inadequacy to tackle power asymmetry.

Lastly, all the three cases uncover the crucial need for oversight authorities to be conferred with enough resources and expertise in order to effectively demand transparency from private companies and perform forward-looking balancing exercises when it comes to Governments and public authorities.

The opportunities deriving from technology's augmented capabilities, especially if employed to serve the public interest, are considerable. However, before welcoming the effectiveness, preciseness and cost-efficiency brought about by ADM, European countries need to improve their preparedness for limiting and managing its potential pitfalls, both from a regulatory and from a governance perspective.

Undoubtedly, the road to a complete understanding and an appropriate handling of algorithmic decision-making is long. However, this journey in Europe seems to have officially started.

# Bibliography


AccessNow, 2018: Human Rights In The Age Of Artificial Intelligence. Report, November 2018. Available at: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

AI HLEG, 2019: Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intelligence of the European Commission, Brussels, 8 April 2019.

Albert, A., 2018: Wie faires Scoring aussehen könnte [How fair scoring could look like]. *Spiegel Online*, 1 December 2018. Available at: https://www.spiegel.de/wirtschaft/service/kreditwuerdigkeit-wie-faires-scoring-aussehen-koennte-a-1241323.html

algo:aware, 2018: State-of-the-Art Report on Algorithmic decision-making. Report, December 2018. Available at: https://www.algoaware.eu/state-of-the-art-report/

AlgorithmWatch, 2019: AlgorithmWatch website. Available at: https://algorithmwatch.org/en/what-we-do/

AlgorithmWatch, 2019: Automating Society - Taking Stock of Automated Decision Making in the EU. Berlin, Matthias Spielkamp, January 2019.

Alpaydin, E., 2014: Introduction to Machine Learning. MIT Press, Cambridge (MA).

Angwin, J., Larson, J., Mattu, S. & Kirchner, L., 2016: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, 23 May 2016. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bendixen, M., 2018: Denmark's 'anti-ghetto' laws are a betrayal of our tolerant values, *The Guardian*, 10 July 2018. Available at: https://www.theguardian.com/commentisfree/2018/jul/10/denmark-ghetto-laws-niqab-circumcision-islamophobic

BEUC, 2018: Automated Decision Making And Artificial Intelligence - A Consumer Perspective. Report, 20 June 2018, Brussels. Available at: https://www.beuc.eu/publications/beuc-x-2018-058_automated_decision_making_and_artificial_intelligence.pdf

Bishop, C. M., 2006: Pattern Recognition and Machine Learning. Springer Science+Business Media, New York.

Bitkom, 2019: Paper Recommendations for the implementation of the Artificial Intelligence Strategy of the Federal Government. Report, 15 February 2019. Available at: https://www.bitkom.org/Bitkom/Publikationen/Handlungsempfehlungen-zur-Umsetzung-der-Strategie-Kuenstliche-Intelligenz-der-Bundesregierung

Bolukbasi, T. et al., 2016: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, Volume 29, pp. 4349-4357.

Bornstein, A. M., 2017: Are Algorithms Building the New Infrastructure of Racism?. *Nautilus,* 21 December 2017. Available at: http://nautil.us/issue/55/trust/are-algorithms-building-the-new-infrastructure-of-racism

Bucher, B., 2018: WhatsApp, WeChat and Facebook Messenger Apps – Global Messenger Usage, Penetration and Statistics. *MessengerPeople*, 3 May 2018. Available at: https://www.messengerpeople.com/global-messenger-usage-statistics/

Buolamwini, J., 2018: Gender Shades. YouTube Video, MIT Media Lab, 9 February 2018. Available at: https://www.youtube.com/watch?time_continue=299&v=TWWsW1w-BVo

Buolamwini, J. & Gebru, T., 2018: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, Volume 81, pp. 1-15.

Burkov, A., 2019: The Hundred-Page Machine Learning Book. Andriy Burkov, Quebec City.

Caliskan, A., Bryson, J. J. & Narayanan, A., 2017: Semantics derived automatically from language corpora contain human-like biases. *Science*, 14 April 2017, pp. 183-186.

Chiappa, S. & Gillam, T. P. S., 2018: Path-Specific Counterfactual Fairness. *arXiv:1802.08139v1*, 22 February 2018. Available at: https://arxiv.org/pdf/1802.08139.pdf

Chouldechova, A., 2017: Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, Volume 5(2), pp. 153-163.

Chouldechova, A. et al., 2018: A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research*, Volume 81, p. 1–15.

Danish Government, 2018: Regeringen vil gøre op med parallelsamfund [The government will do away with parallel communities]. Danish Government website, 1 March 2018. Available at: https://www.regeringen.dk/nyheder/ghettoudspil/

Danish Ministry of Finance, 2018: World-class Digital Service. Report, October 2018. Available at: https://en.digst.dk/media/18772/world-class-digital-service.pdf

Dastin, J., 2018: Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 10 October 2018. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

DataEthics.eu, 2019: DataEthics.eu website. Available at: https://dataethics.eu/

Delker, J., 2019: Europe's silver bullet in global AI battle: Ethics. *POLITICO*, 17 March 2019. Available at: https://www.politico.eu/article/europe-silver-bullet-global-ai-battle-ethics/

Dieterich, W., Mendoza, C. & Brennan, T., 2016: COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Northpointe Inc, 8 July 2016. Available at: http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Digital Minds, 2019: Digital Minds website. Available at: https://www.digitalminds.fi/

Domingo, P., 2015: The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books, New York.

Dzindolet, M. T. et al., 2003: The role of trust in automation reliance. *International Journal of Human-Computer Studies*, Volume 58(6), pp. 697-718.

Edelman, 2019: 2019 Edelman Trust Barometer - Trust in Technology. Report, Edelman, 2019. Available at: https://www.edelman.com/sites/g/files/aatuss191/files/2019-04/2019_Edelman_Trust_Barometer_Technology_Report_0.pdf.

EDRi, 2019: EDRi website. Available at: https://edri.org/

EESC, 2017: Opinion on AI from the European Economic and Social Committee. European Economic and Social Committee, 31 May 2017. Available at: https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence

Ensign, D. et al., 2018: Runaway Feedback Loops in Predictive Policing. *Proceedings of Machine Learning Research*, Volume 81, pp. 1-12.

EU signatories to the Open Letter on Artificial Intelligence and Robotics, 2017: Open Letter on Artificial Intelligence and Robotics. Available at: http://www.robotics-openletter.eu/

Eubanks, V., 2018: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press, New York.

EurAI, 2019: EurAI website. Available at: https://www.eurai.org/

euRobotics, 2019: euRobotics website. Available at: https://www.eu-robotics.net/

European Commission, 2018a: Artificial Intelligence for Europe. COM(2018) 237 final, Brussels, 25 April 2018. Available at: https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

European Commission, 2018b: Coordinated Plan on Artificial Intelligence. COM(2018) 795 final, Brussels, 7 December 2018. Available at: https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence

European Commission, 2018c: Commission appoints expert group on AI and launches the European AI Alliance. Press release, 14 June 2018. Available at: https://ec.europa.eu/digital-single-market/en/news/commission-appoints-expert-group-ai-and-launches-european-ai-alliance

European Parliament, 2017: European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics. (2015/2103(INL)), 16 Febrauary 2017. Available at: http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html#title1

FCAI, 2019: Finnish Center for Artificial Intelligence website. Available at: https://fcai.fi/

Finnish Ministry of Economic Affairs, 2017: Finland's Age of Artificial Intelligence: Turning Finland into a leading country in the application of artificial intelligence. Report, 18 December 2017. Available at: http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf?sequence=1&isAllowed=y

Gadd, S., 2018: Data security blunder in Gladsaxe Municipality compounded. *CPH Post Online*, 14 December 2018. Available at: http://cphpost.dk/news/data-security-blunder-in-gladsaxe-municipality-compounded.html

GDPR, 2016: Regulation (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 Directive 95/46/EC (General Data Protection Regulation) OJ L 119, 4.5.2016, p. 1–88.

German Federal Government, 2018: National AI Strategy. Report, November 2018. Available at: https://www.ki-strategie-deutschland.de/home.html

German Informatics Society, 2018: Study - Regulation of algorithmic decision-making systems. Report, October 2018. Available at: https://gi.de/fileadmin/GI/Allgemein/PDF/GI_Studie_Algorithmenregulierung.pdf

Goodfellow, I., Bengio, Y. & Courville, A., 2015. Deep Learning. MIT Press, Cambridge (MA).

Hasselbalch, G. & Tranberg, P., 2016: Data Ethics - The New Competitive Advantage. Publishare ApS & Spintype.com, Copenhagen.

Hendricks, L. A. et al., 2016: Generating Visual Explanations. *European Conference on Computer Vision 2016* (ECCV 2016), 28 March, pp. 3-19.

Herlocker, J. L., Konstan, J. A. & Riedl, J., 2000: Explaining collaborative filtering recommendations. *Conference on Computer Supported Cooperative Work* (CSCW).

Hunton Andrews Kurth LLP, 2014: Federal German Court Rules on Credit Scoring and Data Subject Access Rights. *Privacy and Information Security Law Blog*, 29 January 2014. Available at: https://www.huntonprivacyblog.com/2014/01/29/federal-german-court-rules-credit-scoring-data-subject-access-rights/

Justitia, 2019: Justitia website. Available at: http://justitia-int.org/en/

Kissinger, H. A., 2018: How the Enlightenment Ends. *The Atlantic*, June 2018. Available at: https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/

Kjær, J. S., 2018a: Småbørns trivsel registreres bag om ryggen på forældrene [The well-being of small children is recorded behind the back of the parents]. *Politiken*, 11 January 2018. Available at: https://politiken.dk/indland/art6288856/Sm%C3%A5b%C3%B8rns-trivsel-registreres-bag-om-ryggen-p%C3%A5-for%C3%A6ldrene

Kjær, J. S., 2018b: Regeringen har lagt sin plan om overvågning af børnefamilier i skuffen [The government has put its plan on monitoring families with children in the drawer]. *Politiken*, 14 December 2018. Available at: https://politiken.dk/indland/art6919255/Regeringen-har-lagt-sin-plan-om-overv%C3%A5gning-af-b%C3%B8rnefamilier-i-skuffen

Kleinberg, J., Mullainathan, S. & Raghavan, M., 2016: Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807v2*, 17 November 2016. Available at: https://njoselson.github.io/pdfs/1609.05807v2.pdf

Kusner, M. J., Loftus, J. R., Russell, C. & Silva, R., 2017: Counterfactual Fairness. *Advances in Neural Information Processing Systems 30*, Volume 30, p. 4066–4076.

Lakkaraju, H. et al., 2017. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. I*nternational Conference on Knowledge Discovery & Data Mining* (KDD), 13 August 2017, pp. 275-284.

Lernende Systeme, 2017: Lernende Systeme website. Available at: https://www.plattform-lernende-systeme.de/about-the-platform.html

Loftus, J. R., Russell, C., Kusner, M. J. & Silva, R., 2018: Causal Reasoning for Algorithmic Fairness. arXiv:1805.05859v1, 15 May 2018. Available at: https://arxiv.org/pdf/1805.05859.pdf

Mølsted, H., 2018. Databeskyttelseslov: Bagdøre i Papes 'oplysningspligt' til borgerne [Data Protection Act: Backdoors in Pape's duty to provide information to citizens]. Version2, 17 May 2018. Available at: https://www.version2.dk/artikel/databeskyttelseslov-bagdoere-papes-oplysningspligt-borgerne-1085164

McCarthy, J., 2007: What is Artificial Intelligence?. Personal website, 12 November 2017. Available at: https//:formal.stanford.edu/jmc/ index.html

Mchangama, J. & Liu, H.-Y., 2018: The Welfare State Is Committing Suicide by Artificial Intelligence. *Foreign Policy*, 25 December 2018. Available at: https://foreignpolicy.com/2018/12/25/the-welfare-state-is-committing-suicide-by-artificial-intelligence/

Miller, A. P., 2018: Want Less-Biased Decisions? Use Algorithms. Harvard Business Review, 26 July 2018. Available at: https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

MyData, 2019: MyData Finland website. Available at: https://mydata.org/finland/

Pasquale, F., 2015: The Black Box Society. The Secret Algorithms That Control Money and Information. Harvard University Press, Cambridge (MA)

Privacy International, 2017. Data Is Power: Profiling and Automated Decision-Making in GDPR. Report, April 2017. Available at: https://privacyinternational.org/sites/default/files/2018-04/Data%20Is%20Power-Profiling%20and%20Automated%20Decision-Making%20in%20GDPR.pdf

Purdy, M. & Daugherty, P., 2018. Why Artificial Intelligence is the future of growth. *Accenture*, 2017. Available at: https://www.accenture.com/t20170524T055435__w__/ca-en/_acnmedia/PDF-52/Accenture-Why-AI-is-the-Future-of-Growth.pdf

Quintarelli, S., 2019: Interview with Stefano Quintarelli, Italian IT specialist and Member of the AI HLEG. (9 April 2019).

Renda, A., 2019a: Europe's Quest For Ethics In Artificial Intelligence. *Forbes,* 11 April 2019. Available at: https://www.forbes.com/sites/washingtonbytes/2019/04/11/europes-quest-for-ethics-in-artificial-intelligence/?fbclid=IwAR00KuLv0IznczQf3EorS8wyO6VAz4_vyUrj0FLS4BeeFCcPHzAv5dkgq80#15b3f5e37bf9

Renda, A., 2019b: Europe's big tech contradiction. *Centre for European Policy Studies*, 2 April 2019. Available at: https://www.ceps.eu/publications/europe%E2%80%99s-big-tech-contradiction

Ribeiro, M. T., Singh, S. & Guestrin, C., 2016: "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *International Conference on Knowledge Discovery & Data Mining* (KDD), 9 August 2016.

Rota, G.-C., 1985: The Barrier of Meaning. *Letters in Mathematical Physics*, Volume 10, pp. 97-99.

Ruckenstein, M., 2019: Phone interview with Minna Ruckenstein, associate professor at the Consumer Society Research Centre and the Helsinki Center for Digital Humanities at University of Helsinki. (17 April 2019).

Scharre, P., 2019: Killer Apps: The Real Dangers of an AI Arms Race. *Foreign Affairs*, May/June 2019. Available at: https://www.foreignaffairs.com/articles/2019-04-16/killer-apps?utm_campaign=reg_conf_email&utm_medium=newsletters&utm_source=fa_registration

Shutton, R. S. & Barto, A. G., 2017: Reinforcement Learning: An Introduction. MIT Press, Cambridge (MA).

Signatories of the Declaration of cooperation on Artificial Intelligence, 2018: Declaration of cooperation on Artificial Intelligence. 10 April 2018. Available at: https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence

SIRI Commission, 2018: Ethics and AI scenarios from the SIRI Commission. Report, 12 September 2018. Available at: https://ida.dk/om-ida/temaer/siri-kommissionen/etik-og-ai-scenarier-fra-siri-kommissionen

SPIEGEL Data & BR Data, 2018: OpenSCHUFA, *Spiegel Online*, November 2018. Available at: https://www.spiegel.de/wirtschaft/service/schufa-so-funktioniert-deutschlands-einflussreichste-auskunftei-a-1239214.html

Spielkamp, M., 2019: Phone interview with Matthias Spielkamp, founder and executive director of the non-profit organisation AlgorithmWatch. (12 April 2019).

Spielkamp, M. & Kaiser-Bril, N., 2019: Resist the robot takeover. *POLITICO,* 12 February 2019. Available at: https://www.politico.eu/article/resist-robot-takeover-artificial-intelligence-digital-minds-email-tool/

Teach, R. L. & Shortliffe, E. H., 1981: An analysis of physician attitudes regarding computer-based clinical consultation systems. In: Anderson J.G., Jay S.J. (eds) *Use and impact of computers in clinical medicine*, Springer, New York.

Tranberg, P., 2019: Interview with Pernille Tranberg, co-founder of the Danish think tank DataEthics.ru and independent advisor in data ethics for companies, authorities and organisations. (17 April 2019).

Wachter, S., Mittelstadt, B. & Floridi, L., 2017: Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, Volume 7(2), p. 76–99.

Wang, D., 2001: Unsupervised Learning: Foundations of Neural Computation: A Review. *AI Magazine*, Volume 22(2), pp. 101-102.

WP29, 2018: Guidelines on Automated individual decision-making and Profiling
for the purposes of Regulation 2016/679 by the Article 29 Data Protection Working Party, adopted on 3 October 2017, as last Revised and adopted on 6 February 2018. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053

Yona, G., 2017: A Gentle Introduction to the Discussion on Algorithmic Fairness. *Towards Data Science*, 5 October 2017. Available at: https://towardsdatascience.com/a-gentle-introduction-to-the-discussion-on-algorithmic-fairness-740bbb469b6

Zuboff, S., 2019: The age of surveillance capitalism. Profile Books Ltd, London.