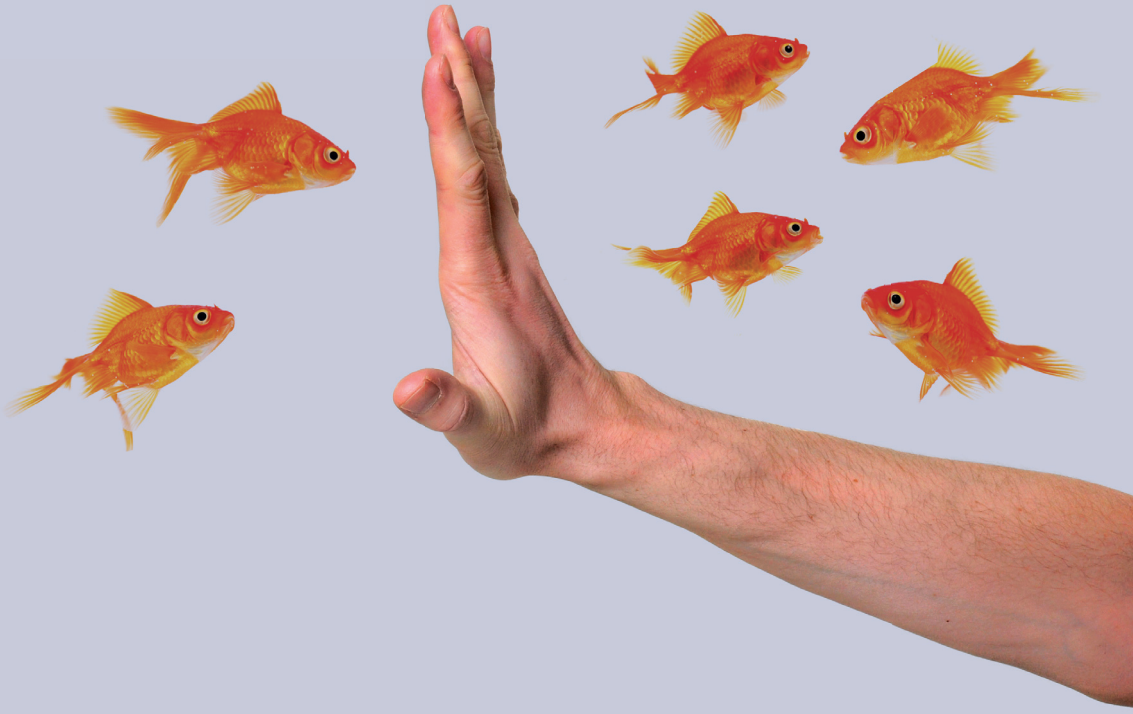
INSTITUT
MONTAIGNE

Algorithms:
Please Mind the Bias!



REPORT MARCH 2020

INSTITUT
MONTAIGNE

Algorithms: Please Mind the Bias!

REPORT – MARCH 2020

There is no desire more natural
than the desire for knowledge

TABLE OF CONTENTS

FINDINGS AND CHALLENGES

Despite the risk of bias in some cases, algorithms in many ways actually represent an advance where discrimination is concerned. Men and women are often consciously or unconsciously biased, inconsistent in their decisions. **Using an algorithm means formalizing rules that apply to everyone, measuring the results, and ensuring that no bias exists.**

Algorithmic biases leading to discrimination are rarely due to an incorrect code in the algorithm. Incomplete or poor-quality data, or data that reflect biases in society, are much more often at the root of such biases. The fight against algorithmic bias is therefore above all a fight against discrimination that already exists on a daily basis. The challenge is not only to produce algorithms that are fair, but also to reduce discrimination in society.

This battle is difficult for numerous reasons. First of all, **it is not simple to define a bias-free algorithm.** While some biases are voluntary, such as promoting need-based scholarship recipients in the school admissions, others are involuntary or ignored, leading to discrimination against certain groups.

A fair algorithm, i.e. one that treats users fairly, is close to an algorithm without bias. However, this can never be fully guaranteed. Taking equity into account does not help designing unbiased algorithms, since equity can take different forms. Assessing what is fair involves an inherent cultural dimension and depends on each situation. The ethical attitude will not be the same in the case of one algorithm that analyzes lung X-rays and another that recommends political advertisements. In addition, total fairness between individuals and complete fairness between groups are fundamentally incompatible. **There will always be societal and political choices to be made.**

Then, **correcting an algorithm to make it fair often means reducing its performance with respect to its initial design criteria.** When we develop an algorithm, we choose one or many metrics allowing to optimize it and assess whether it reaches its goal. Adding constraints means limiting the capacity to optimize the algorithm vis-à-vis its initial performance criteria. It is always more difficult to pursue many goals at the same time rather than a unique one. It will therefore be difficult and costly for many actors to combat discriminations caused by algorithms.

Finally, **combating algorithmic bias means achieving an equilibrium between protecting citizens against discrimination on the one hand, and giving the possibility to experiment, crucial to the digital economy, on the other.** Restricting the use of algorithms, on suspicion of biases means depriving ourselves of new tools that could make our decisions more objective. It means curbing the growth of the French digital industry and accepting American and Chinese technological superiority in the long term. Adopting a *laissez-faire* approach would mean ignoring the destructive potential of such innovations for our social fabric.

RECOMMENDATIONS

Faced with these issues, we must be clear: we recommend neither a law against algorithmic bias common to all sectors of activity, nor a systematic check by the State of the absence of bias in algorithms.

Numerous texts that deal with discrimination already exist. They apply to both the physical and digital worlds and are likely to limit the risk of bias insert period. In view of society's limited hindsight in this field, a specific law on algorithmic bias would risk inhibiting innovation without actually solving the underlying problem.

The General Data Protection Regulation (GDPR) has shown that the use of personal data is far too widespread for a public agency to be able to verify all such data before they are used. We believe that the same will be true for algorithms and that it is illusory to expect the State to check each and every algorithm to ensure that they are ethical before they are implemented.

We have endeavored to formulate recommendations that are as realistic as possible in order to allow the rapid development of new technologies within a framework that respects our lifestyles.

Test the presence of bias in algorithms in the same way that the side effects of medication are tested

Like new drugs, it is complicated to understand how all algorithms work, especially those based on artificial intelligence. Furthermore, understanding how they work does not guarantee that they will be bias-free. It is ultimately by testing for the absence of bias that we will create confidence in the fairness of algorithms.

Testing the fairness of an algorithm has a cost and requires test data that specifically include some sensitive information (gender, social origin). Algorithm developers and purchasers will need to incorporate this constraint, and implement functional or performance testing to ensure the absence of bias. In some cases where the creation of these databases is difficult or problematic, the State could be responsible for their compilation.

Promote active fairness, rather than hoping for fairness by ignoring diversity

In order to combat discrimination, France has long chosen to blindfold itself, to see nothing of individuals beyond their status as citizens. As far as algorithms are concerned, this approach is no longer sufficient. An algorithm can introduce biases against women, even if the gender of the variables used has been explicitly excluded: it is easy to "guess" gender from other information such as buying women's products. To combat discrimination, it must therefore first be possible to detect it.

We need to move from an approach that hopes for fairness through unawareness to one of active fairness. We need to accept that the fairness of an algorithm is not achieved by excluding all protected variables such as gender, age or religion. On the contrary, it is obtained by including them, and by testing the independent nature of the result with respect to these variables. To achieve this, it is necessary to have access to this protected information. But if this information is protected, it is precisely because it can be a source of discriminations. The collection and use of this sensitive data must therefore be strictly supervised, and be limited uniquely to the purpose of testing and restricted to a sample of the users concerned. Moreover, such an approach would have to be the subject of an impact study declared beforehand to the CNIL (the French data protection authority).

Require greater stringency for high-risk algorithms (fundamental rights, security, access to essential services)

The sensitivity of an algorithm with respect to society obviously depends on its sector of activity, but above all on its potential impact on citizens. This impact is significant when the algorithm can restrict access to essential services such as a bank account or a job, endanger security (health, police), or violate fundamental human rights. These areas are already subject to strong discrimination obligations. When an algorithm is introduced in these areas, it cannot be at the cost of lowering requirements.

For these algorithms, we recommend an ad hoc framework integrating transparency obligations with regard to the data used and the objectives set for the algorithm, as well as a right of appeal against the decision taken. The creation of such a framework does not require a new law on algorithmic bias, but rather the implementation of good practices in companies and administrations, the use of existing legal provisions, and the addition of provisions in sectoral legislation on a case-by-case basis.

Ensure team diversity within algorithm design and deployment projects

Algorithms are transforming the business model of companies. Managers and users must therefore be increasingly involved in their design.

Making a decision on what is fair behavior for a given algorithm and its main parameters carries significant societal and economic impact and cannot be the responsibility of technical experts alone. It is more than ever necessary to ensure a multilateral approach in taking decisions of this nature. Beyond professional diversity, it is now clear that socially diverse teams are better suited to prevent biases and avoid reproducing discriminations.

In order to prevent algorithmic bias, algorithm design, production and steering teams must include as the standard a diversity of profiles, skills, experiences, ages and gender.

And to take things further...

Beyond these four strong recommendations that we are putting forward, we are convinced that a great deal of work remains to be done in terms of training. This concerns researchers and developers, of course, especially in the area of algorithmic bias, but also leaders and citizens within the more general framework of artificial intelligence, so that everyone can take ownership of both the opportunities associated with this technology and its inherent risks.

This vigilance should also be reinforced in organizations implementing charters and best practices. These initiatives, which we noted during our interviews, must be encouraged as, together with technical and operational measures, they would make it possible to generate collective awareness around the dangers of algorithmic bias.

Finally, vigilance must be external and, in the case of high-risk algorithms, it would seem judicious to strengthen controls. This could be done firstly via the issuing of labels, the emergence of which should be supported. Such labels would guarantee the quality of the data used and of the organizations developing the algorithms, the existence of control procedures, and the auditability of these algorithms. The industrial sector would notably need such guarantees in order to take full advantage of the algorithmic revolution. Furthermore, with respect to high-risk algorithms, the ability to audit and monitor certain requirements could be entrusted to a third party or to the State.

INTRODUCTION

An investigation conducted in 2016 revealed that the COMPAS software designed to predict the potential risk of recidivism among criminals, and used daily by American judges to decide whether or not to grant bail, discriminated against African-American populations. In early 2019, researchers accused Facebook's job recommendation algorithm of promoting women less frequently. In late 2019, an Apple Card algorithm was denounced for discriminating against wives by automatically attributing them much lower credit limits than their husbands. Given these and other controversies, more and more algorithmic biases are being brought to the public's attention in the United States.

This report attempts to provide a French perspective on this issue, today essentially viewed through the lens of American controversies. Indeed, there are few public examples of algorithmic bias in France or in continental Europe, especially bias leading to discrimination with respect to the 25 criteria protected under French law (age, gender, religion, etc.[1]). The rare established cases involve American companies. However, digital technology and artificial intelligence will never be able to develop in France if they are vectors of massive discrimination. Such large-scale automation of unfair decisions would simply be unacceptable for our society.

Our work continues the study by Télécom Paris and Fondation Abeona, *Algorithmes : biais, discrimination et équité* ("Algorithms: Bias, Discrimination and Fairness"), published in 2019. On the basis of this technical observation, and through the forty or so interviews and multiple working group meetings, our aim is to provide concrete solutions to ensure the deployment in France, by all actors concerned, of ethical and useful algorithms.

1 Origin, gender, age, family status, pregnancy, physical appearance, economical situation, patronym, health, loss of autonomy, morals, genetics, sexual orientation, gender, political or philosophical opinions, languages, real or supposed ethnic affiliation, nation or pretended race, residency, bank domiciliation.

I

ALGORITHMS ARE BOTH REMEDIES AGAINST AND CAUSES OF DISCRIMINATION

Artificial intelligence – or rather machine learning – has made tremendous strides in recent years. All areas of our contemporary lives leave digital traces and generate data that can be used to develop algorithms. These help predict defects in mechanical parts, recognize tumors, or propose a bank loan rate.

The last decade saw an unprecedented rise of the tech industry and its services reinforced by artificial intelligence. But despite all of its promises, despite the victory of the geeks, the decade ended with two years of techlash. Concerns about the consequences of artificial intelligence are multiplying: for our work, our media, our democracies.

The almighty AI, surpassing humans in its intelligence, is not for tomorrow. Before fearing it, it is urgent to squarely examine and understand the way algorithms fit into our lives, here and now.

Algorithms are decked out in the finery of neutrality and objectivity, but they are embedded in our societies, with their flaws and structural inequalities. Many examples document how algorithms sort, classify and exclude certain groups[2]. Like digital databases, like their paper ancestors before them, they have the potential to exacerbate certain inequalities.

In our western societies, the spread of algorithms is concomitant with a greater awareness of discrimination and, more broadly, of structural disadvantages. Be it gender, ethnic origin, sexual orientation or age, our intolerance of inequalities between groups is growing.

The study of "algorithmic bias" – the impact of algorithms on discrimination – is largely dominated by Anglo-Saxon voices and cases. Whether it concerns justice, recruitment, credit or facial recognition, the most heated controversies tend to come from the United States[3].

2 See the article by Bertail P., Bounie D., Clémençon S. et Waelbroeck P., *Algorithmes : biais, discrimination et équité*, 2019.
3 Involving COMPAS, Amazon, Apple Card or Microsoft, respectively.

Does a similar situation exist here in France? Algorithms are not developed in the same way here. We have stricter rules governing the collection of personal data and certain practices are formally prohibited, such as gender segmentation in the insurance sector.

We are convinced that algorithms can bring transparency, assessment, and objectivity to our systems, which are too often undermined by discrimination. Through our interviews, we sought to identify the risks posed by algorithms with respect to discrimination. We then tried to draw up a few recommendations so that fears – sometimes exaggerated, but also sometimes legitimate – could be overcome, and so that France can progress towards "trustworthy AI."

A. Algorithms are sometimes a useful remedy against discrimination

Organizations in many sectors are trying to deploy algorithms to improve their decision-making, objectify it, and in some cases prevent discrimination:

Recruitment: The world leaders in temporary employment are implementing algorithms to propose to their recruitment consultants those profiles that seem best suited to a job offer. Algorithms generally integrate the skills of the candidates and the satisfaction of their past employers. They are able to detach themselves from the "standard profiles" that employers generally turn to and that lead them to systematically recruit the same types of profiles. In addition to opening up the horizons of teams, algorithms also make it possible to reduce hiring discrimination. It has long been known that being called Rachid or Mariam has an impact on the chances of being recruited, and that an anonymous résumé does not change all that much[4]. We also know that when certain prerequisites are necessary (several internships in the industrial sector, for example), candidates who come from disadvantaged categories are often excluded. Algorithms that are based on skills, and no longer experience only, can overcome these problems.

Housing: Renting a place to live is sometimes a lengthy and complicated process, especially in the tight markets of big cities. For some, the process can be even longer and more complicated. A person of North African origin who states that he/she is a civil servant, and thus benefits from job security, will receive an answer to his/

4 Behagel L., Crepou B., Le Barbauchon T., *Evaluation de l'impact du CV anonyme*, PSE ; voir aussi, Institut Montaigne, *Discriminations religieuse à l'embauche, une réalité*, October 2015.

her rental application in 15.5% of cases, against 42.9% of cases for an applicant of French origin with the same professional profile[5]. Certain measures can be effective in reducing the difference in response rates, such as a reminder of the law by the French Défenseur des Droits (Ombudsman), but their effects diminish after nine months, and disappear completely after 15 months[6]. Algorithms for automatically recommending applicants for housing offers could make it possible to reduce this discrimination by evaluating two people with the same profile in the same way.

Justice system: The American justice system is known for its history of discrimination against African-American people. Some accuse algorithms of reproducing the same discrimination, while others see in them an opportunity to do away with this situation. Bail has been a textbook case since the controversy around the COMPAS algorithm[7]. Currently, criminal risk assessments are performed to assess a suspect's risk of recidivism to help the judge decide whether or not this person should be released on bail. These assessments over-estimate the risk that an African-American suspect will reoffend on bail, and thus lead judges to grant bail to fewer Blacks than Whites. And yet these assessments remain more accurate than the judgment of the police or the judges alone. Algorithms attempt to assess this risk on the basis of past history, and propose this assessment to the judges. By some measures, these algorithmic assessments are always to the disadvantage of African-American offenders. Nevertheless, they are more accurate than criminal risk assessments, and allow judges to be more consistent and precise in their decisions. These algorithms are far from perfect, but their use represents a step in the right direction.

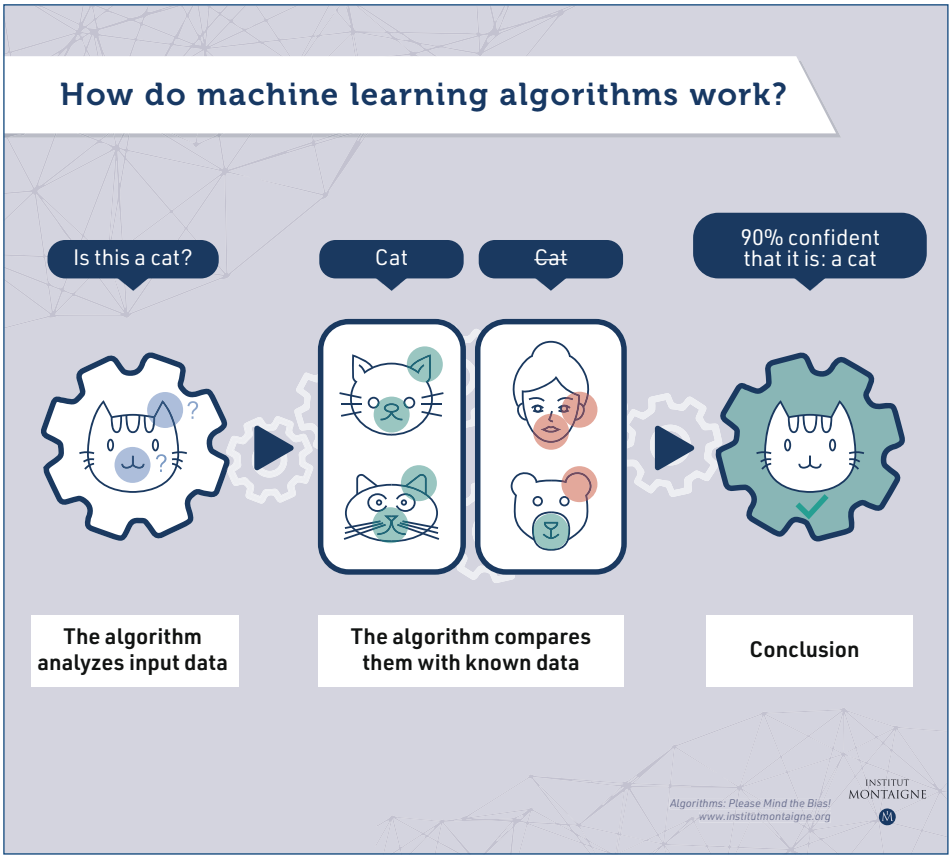
Such applications are not yet possible in France. However, it is known that, all other things being equal, unemployed people are 1.8 times more likely to be tried immediately than people in stable employment[8]. For some judges, such decisions are logical and deliberate (to avoid making it impossible for the justice system to later find and judge offenders). However, algorithms would make these choices explicit, and reduce the remaining randomness in judges' decisions.

5 Bunel M., l'Horty Y., Du Parquet L., Petit P., *Les discriminations dans l'accès au logement à Paris ; une expérience contrôlée*, ffhalshs-01521995f, May 2017.

6 French Défenseur des droits (Ombudsman), *Test de discrimination dans l'accès au logement selon l'origine*, October 2019.

7 Goel S., Shroff R., L. Skeem J., Slobogin C., *The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment SSRN*, December 2018.

8 Gautron V., Retière J-N, *La justice pénale est-elle discriminatoire? Une étude empirique des pratiques décisionnelles dans cinq tribunaux correctionnels*, Colloque « Discriminations : état de la recherche », Alliance de Recherche sur les Discriminations (ARDIS), Université Paris Est Marne-la-Vallée, France. ffhalshs-01075666, December 2013.



What is AI?

An algorithm is a sequence of operations or instructions that allow for obtaining a result. In many ways, a recipe is an algorithm. Algorithms exist in many forms. One of these forms, machine learning, has undergone significant development in recent years, partly due to the growth in the quantity of data available.

.../...

How does a machine learning algorithm work?

An illustrative example is that of pattern recognition algorithms. In this type of problem, the algorithm must accomplish the following task: using input data X, it must automatically recognize the category Y associated with each object/individual X, with minimal risk of error. Category Y is of a given type, specified in advance.

A wide range of applications correspond to this formulation, from biometrics and credit risk management to assisted medical diagnosis and prognosis. In the case of computer vision, for example, X will correspond to a pixelated image and the output Y to a "label" associated with the image indicating the possible presence of a specific object in the image (tumor, cracked rib).

The rule for deciding which "label" to attach to each image is determined by a learning algorithm. This algorithm operates on a database of already labeled data (i.e. the "learning data") which already associate images with the correct label.

The importance of large databases for machine learning

The objective of a learning algorithm is to discover in the learning data the regularities that will allow it to find the label associated with new data not yet observed. Faced with a new X-ray image it has never seen before, the algorithm must be able to determine with sufficient prediction the presence of a cracked rib. The goal is to minimize the probability of taking random X data and mislabeling it as Y.

To be effective when faced with new data, an algorithm must have learned from a wide variety of cases. Hence the importance of a large quantity of data and significant computational capabilities. The megadata of the web, the huge libraries of images, sounds or texts that are "labeled" – often by humans – are therefore crucial.

B. Algorithms pose new risks for discrimination, but the French debate remains influenced by American examples

a. Should we fear algorithms?

The launch of the iPhone in 2007 ushered in a period of technological optimism concerning the digital revolution. A decade later, the situation is one of techlash. Not a month goes by without a new investigation revealing invasive practices in our private lives, or algorithms used for dubious purposes. American mathematician Cathy O'Neil has qualified algorithms and big data as "weapons of math destruction"[9]: tools that, under the guise of objective mathematical formulas, reinforce inequalities and discrimination, thereby amplifying the effects of inequalities.

In 2015, Amazon implemented an algorithm to facilitate the recruitment of talent[10]. Learning from hundreds of thousands of résumés received by Amazon over a ten-year period, the algorithm assigned a rating ranging from 1 to 5 stars. However, the use of this algorithm was quickly suspended due to its inability to select the best candidates and its bias against women. Indeed, it frequently assigned poor scores to qualified female profiles, and systematically proposed under-qualified male candidates. It disadvantaged résumés containing the words "women's", including "women's chess club captain", and favored résumés containing the words "executed" or "captured", which are more common in men's résumés. It was the quality of the learning data that was questioned, with men constituting the overwhelming majority of managers recruited in the past, while many women were over-qualified for their positions. As a result, the algorithm had learned to underestimate the résumés of women.

Despite the benefits that algorithms could bring to the fight against discrimination, it is mistrust and fear of their effects that dominate, at least in the public debate.

The exclusion or unfavorable treatment of certain groups already existed long before the advent of algorithms. Racism on credit and sexism in hiring did not wait until the third millennium. Nor did algorithms come into being only in 2010. Predictive police testing preceded the rise of machine learning, while the banking and insurance sectors have long incorporated statistical models and algorithms into their operating methods.

9 O'Neil C., *Weapons of Math Destruction*, Penguin Books, June 2017.
10 Reuters, *Amazon scraps secret AI recruiting tool that showed bias against women*, October 2018.

Whether they reduce, maintain or aggravate discrimination, algorithms are systematically condemned. There are several reasons for this.

Where our ideals are concerned, an algorithm that would reduce discrimination but without making it disappear altogether would probably be considered insufficient. And the more we make progress towards equality, the more inequality seems intolerable to us. This is Tocqueville's paradox:

> *"In democratic peoples, men easily obtain a certain equality; they cannot attain the equality they desire. It retreats before them daily but without ever evading their regard, and when it withdraws, it attracts them in pursuit. They constantly believe they are going to seize it, and it constantly escapes their grasp. They see it from near enough to know its charms, they do not approach it close enough to enjoy it, and they die before having fully savored its sweetness."*
>
> *Alexis de Tocqueville, Democracy in America*

Even assuming that algorithms are but neutral tools, they can still perpetuate or consolidate previous discrimination. They can also accurately measure a bias that may have been known, but the magnitude of which had remained hidden. Each new algorithm is an opportunity to clarify the inequalities inherent in our societies.

Finally, learning algorithms do have the potential to worsen the situation with regard to discrimination:

- The large-scale implementation of these algorithms simultaneously accompanies a digitization of our transactions and behaviors that generates data in all areas of our societies. The fields of application extend far beyond banking, insurance and major consumer brands.
- Machine learning algorithms appear as black boxes, the operation of which sometimes remains inscrutable even for their designers. They are more complex than their predecessors. In the era of big data, an algorithm can be supplied with dozens of variables and tasked with finding the most appropriate combinations and weightings. Nevertheless, there is not always an intelligible way to "understand" the given result.
- When algorithms merely reproduce existing inequalities or types of discrimination, they can generalize previously circumscribed discrimination. The COMPAS algorithm learns from a single forensic dataset, that of Broward County in Florida. A judge in Oregon who uses COMPAS can thus recover the historical biases of the Broward County court system. The massive use of such an algorithm in the United States could multiply the effects of discrimination limited to a single county in Florida.

b. Reconstructing the French debate on the impact of algorithms

In the media as well as in academic research, the omnipresence of Anglo-Saxon examples, particularly American, is striking. The deployment of algorithms in the United States is much more advanced than in France as a result of regulations that are less strict and a more developed digital industry. Indeed, some of the most controversial applications in the United States could simply not have happened in France. Article 10 of the 1978 French Data Protection Act states that "No court decision involving the assessment of an individual's behavior may be based on an automatic processing of personal data intended to assess some aspects of their personality."

The discrepancy between France (and, more broadly, continental Europe) and the United States is both good and bad news. We have the opportunity to imagine a supervisory framework that would avoid the excesses of certain systems deployed on the other side of the Atlantic, and that would correspond to our system of values. But we are also lagging behind in mastering these tools. There is thus a great risk of further hindering the digital transformation of the French economy due to a lack of understanding of the issues at stake.

Although much fewer organizations in France deploy machine learning algorithms than in the United States, the French use the digital services of GAFA on a daily basis, the functionalities of which are enhanced by algorithms.

According to the IFOP (French Institute of Public Opinion), 80% of French people consider that the presence of algorithms in their lives is already huge, but a little more than half of them admit that they do not exactly know what algorithms are[11]. Apart from a few areas (banking and insurance, advertising targeting, GAFA's purely digital services), and beyond the effects of advertising, decisions assisted or made by algorithms remain few and far between in France.

11 CNIL, *Éthique et numérique : les algorithmes en débat*, January 2017.

II

ALGORITHMIC BIAS: AN OLD AND COMPLEX PROBLEM

A. Biases pre-exist algorithms and are mainly found in the data they use

Machine learning algorithms learn and decide from data produced by humans and converted into digital formats. This can be relational data from social networks, purchasing behavior used for marketing purposes, music preferences from streaming platforms, videos and photos posted on the Internet, text messages exchanged, Google search histories, recruitment or credit decisions, and so on.

These data represent a digital mirror of human behavior. They directly and accurately reflect our habits and therefore our biases. With the development of sensors and the Internet of Things, more and more detailed data are being collected. If we are not careful, algorithms that learn from these data and their biases will have the capacity to standardize and amplify discrimination.

Discrimination and bias, what are we talking about?

In technical parlance, algorithmic bias is the mean deviation between the prediction made and the value that one was trying to predict. In concrete terms, this can be the deviation between the number of X-ray images labeled by the algorithm as "comprising a tumor" and the number of X-ray images that actually do comprise a tumor. High bias means that the algorithm lacks relevant relationships between input and output data to make correct predictions.

These **"technical biases"** are taken into account by engineers, as they directly reduce the performance of the algorithm. They become part of the public discussion about "algorithmic bias" when they disadvantage a specific group. This is not systematic: an algorithm that recognizes oak trees less often than birch trees on images is technically biased, but without prejudice in terms of discrimination.

.../...

Algorithmic biases go beyond technical biases. The algorithm can be very accurate technically while being "biased" from a social point of view. These biases are condemned as discrimination because they often select and arbitrate to the disadvantage of already disadvantaged populations. These **"societal biases"** are in fact the reproduction via the algorithm of biases already present in society.

The distinction between technical bias and societal bias is important because **these biases do not attract the same attention on the part of developers.** Technical biases reduce the performance of the algorithm, hindering the achievement of its objective. Reducing technical biases has a cost, but often also a clear benefit for developers.

Conversely, societal biases generally reproduce biases already present in society. Following these biases allows algorithms to perform better – at least in a limited sense. When it comes to advertising or job postings, sticking to stereotypes can maximize the number of clicks on the ads.

Algorithmic biases can also be conscious decisions to support a business strategy. Google, for example, was fined € 2.4 billion for favoring its own products in Google Shopping search results, at the expense of those of its competitors. The algorithm was deliberately biased.

For the purposes of this report, the term algorithmic bias refers to both technical biases that are well known to statisticians, as well as societal biases that are less well defined for them.

Technical biases often arise from the data they are trained on

Based on the article by Patrice Bertail, David Bounie, Stéphan Clémençon and Patrick Waelbroeck, Algorithmes: *Bias, Discrimination and Equity*, published by Telecom Paris and Fondation Abeona

Several types of technical biases exist and they could interfere with the algorithm's performance. This could be a methodology flaw in the algorithm development, but more often the problem stems from a data quality or data distribution issue - that is data being representative of the population sample.

Some data is complicated, if not impossible, to collect. Sometimes it is approximated or is replaced with proxies. This is what is known as an **omitted variable bias.** For example, skills such as leadership or emotional intelligence are difficult to measure and may be negatively correlated with academic performance. A candidate-selection algorithm that takes into account only the academic performance would surely fail to identify certain high-potential individuals, even though this was its main goal.

Beyond the choice of variables, **selection bias** can occur when the learning sample is not representative of the population covered, for example, in medicine, when the algorithm is trained on data of predominantly white people[12]. There is a well-known example of a facial recognition algorithm, which worked better on white people than on people of color because the data used for training were images put together by white developers. An evaluation[13] carried out by the American National Institute for Science and Technology on 189 algorithms by 99 developers thus established that the false positive rates (of people wrongly recognized by their smart phone for example) were 10 to 100 times higher for people from Africa and East Asia (except for the latter in the case of an algorithm developed in China).

The most common source of technical bias is certainly the quality of learning data, which can produce database bias. If the labels on the training images are wrong (many images are labelled “sheep” when the image is actually that of a goat), the final result will inevitably be biased, in the technical, statistical sense. This is what happens when the algorithm is trained on so-called “biased” data (e.g. sexist views on employee performance). In another example, the algorithm may appear to function properly when it has only found clues in the training data: instead of finding X-rays that show tumors, this algorithm could actually be trained to find X-rays stored by the data scientist in the C://Images/Photos_tumors folder. As a result the algorithm will not be able to recognize images of real tumors.

We call such bias technical because it stems from technical errors, but it is not without consequences for society. For example, the Shirley card (see the box below), a technical standard created by Kodak and calibrated for white people only, led to the impossibility for black people to be correctly seen on photo images - and thus become invisible, in both literal and figurative senses.

The Shirley cards illustrate that, when it comes to technical bias, the problem often lies on the data side, especially in terms of data quality or representation. If Kodak

12 Buolamwini, J. et Gebru, T. Gender Shades: *Intersectional Accuracy Disparities in Commercial Gender Classification* in Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 2018.
13 NIST, *Face Recognition Vendor Test (FRVT)* Part 3: Demographic Effects, 2019.

developers had chosen to build a base of Shirley cards that was representative of the diversity of skin colors, the system would not have discriminated against non-white profiles. This example also reminds us that technical standards are not benign, a useful reminder when we see today the battle led by Chinese companies to define technical standards for facial recognition algorithms.

Kodak's Shirley cards, from data to technical bias

Bias in emerging technologies existed long before artificial intelligence algorithms. A notable example is Kodak's color film that standardized degraded quality of photographs for people with non-white skin. In the early 1940s, Kodak was one of the few suppliers of color photographs. To calibrate skin color, shadows, and light during photo development and printing process, the company created a Shirley card, named for the first model who was a Kodak employee. This was a portrait of a white-skinned woman with black hair. As the years passed, the Shirley models changed, but the successive images all conformed to the socially acceptable white beauty standards. These cards were distributed in all the laboratories of the world with kits containing original negatives, allowing to calibrate photo equipment. The chemical inputs were optimized to magnify colors with light tones, and thus poorly captured dark tones. Those absorbing more light required different development to provide optimal rendering. Black skin came out poorly visible, with only the teeth and eyes appearing in high contrast.

The American anti-discrimination movements never questioned Kodak. At the time, the dominant assumption was that the problem was scientific, that there were no technical solutions to better represent black skin. Yet in the 1970s, Jean-Luc Godard refused to film in Mozambique with Kodak equipment, accusing it of being racist.

A partial solution came from the corporate world. Two of Kodak's major customers were candy and furniture manufacturers. They complained that advertisements for their chocolates and dark furniture did not look good on Kodak film. The company then developed new films and produced a new Shirley: 3 women of Caucasian, African-American and Asian descent.

For over 30 years, people of color were denied a visual identity, while the controlling white majority retained and protected full control of the normative image through the use of innovative but highly biased technology.

Several societal biases can be encoded in the data used by the algorithms

Psychology distinguishes two types of bias that distort our decisions: **emotional (or affective) bias,** and **cognitive bias** (especially stereotypes). While emotional bias leads us to refuse to believe in unpleasant realities, stereotyping refers to treating a person according to the group to which he or she belongs (and the traits we associate with that group), rather than on his or her individual characteristics.

These two types of bias existed in society well before the advent of algorithms and artificial intelligence. In the 21st century, few stereotypes are openly acknowledged. Only 10% of the population[14] explicitly acknowledges being biased. However, implicit stereotypes are widespread, affecting our judgement and often leading to discrimination.

Anthropologist Edward T. Hall's research has highlighted the link between stereotypes and the way information is exchanged[15]. In societies he describes as "high context", this exchange takes place via implicit elements such as body language or social behavior. These clues are easily grasped by a community member but are more complex for a foreigner to access, thus creating integration barriers. Japan and France are examples of high context cultures. In these societies, traditions are present on a daily basis, and judgement is made through the application of norms. These environments are exposed to the force of stereotypes, which are tools of reading and understanding context.

Algorithms can spread these types of biases in society, either by simply replicating already biased human decisions, or because the developers themselves believe in these stereotypes.

As a result, developers can sometimes allow cognitive bias interfere with the way they design or interpret their models. This may lead them to implement models that correspond to their outlook. Michal Kosinski and Yilun Wang, two Stanford researchers, proposed in 2018 to detect an individual's sexual orientation according to his or her facial morphology[16]. Physiognomy - the idea that physical appearance gives a glimpse of character - has a long and troubled history. Many researchers have pointed out that this algorithm detected individual styles (beard, makeup, glasses) rather than any

14 Fiske S. and Taylor S. *Social Cognition: From Brains to Culture*, 2nd edn, Sage: London,1984.
15 Hall, E. T., *Beyond Culture*, Anchor, Décembre 1976.
16 Kosinski M., Wang Y. *Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images* in Journal of Personality and Social Psychology, February 2018.

"homosexual physiognomy"[17]. Yet these style elements were themselves signals that people send out to demonstrate their belonging to certain groups. Far from detecting sexual orientation on the basis of facial features, and reinforcing a certain theory about sexual orientation, the work of Kosinski and Wang is probably more revealing of their own stereotypes than of any reality.

Algorithms have been repeatedly accused of spreading stereotypes, particularly with regard to gender. For example, women tend to respond only to job offers that they believe they have a high probability of getting, and thus algorithms learn to recommend them certain types of jobs. Another example of stereotype bias is found in the co-occurrence of words: "woman" with "stylist" or "hairdresser", while the word «man» is associated with "captain", "chief", or "financier"[18]. These stereotypes can feed and bias algorithms, especially those that recommend job offers.

Beyond cognitive and affective biases, **economic bias** is another type of bias encountered. An algorithm may contain a bias voluntarily or involuntarily for reasons of business strategy. An algorithm that simply optimizes the cost-effectiveness of a job posting displays fewer advertisements to young women than to young men. This is because advertising space for young women is more expensive than advertising space for young men[19]. Thus, it will be less costly for the algorithm to prefer men for these job ads. The commercial strategy to recruit while minimizing recruitment costs could thus lead to discrimination against women.

One could argue that algorithms are not responsible for societal biases. However, they can multiply and standardize them, which is where the risk comes in. Indeed, systemized societal biases can have profound effects on individuals.

Bias and discrimination could become self-fulfilling prophesies by influencing the behavior of discriminated groups in the direction predicted by stereotypes. A survey in French supermarkets[20] has shown that exposure to managers with significant biases negatively affects work performance of minority staff[21]. When the working hours of minority staff coincided with those of managers with stereotypes, cashiers validated

17 Aguera y Arcas,B. *Do algorithms reveal orientations or just expose our stereotypes?* Medium Equality, January 2018.
18 Bolukbasi T., Chang K-W., Zou J., Saligrama V., Kalai, A. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, July 2016.
19 Lambrecht, A. and Tucker, C. *Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads*. March 2018.
20 Glover D., Pallais A., Pariente W., *Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores*, The Quarterly Journal of Economics, Volume 132, Issue 3, August 2017.
21 Persons with a northern African or sub-Saharan name.

items less quickly, were absent more often and left work earlier, leading to lower wages. Managers expressing stereotypes, however, do not treat minority employees differently, but simply seem to spend less time with them. Thus, stereotypes do not even need to guide the explicit action of managers to have an impact on pay differentials and recruitment habits.

Facing these types of risks, bias impact analysis should not be limited to technical biases: this research must fully integrate the risk for algorithms to transmit various forms of societal bias.

B. A fair algorithm has multiple contradictory definitions; is it the role of organizations to choose which one to apply?

Whether we are talking about technical bias or societal bias, the cultural dimension of algorithmic bias cannot be ignored. Beyond effectiveness, academic research usually defines three qualities with which algorithms can be endowed: fairness, obviously, but also neutrality and loyalty[22]. Each can constitute a partial guarantee against algorithmic bias. These three elements are subject to multiple definitions depending on the context in which the algorithm will be used, but also depending on the cultural environment (American or European). In absolute terms, unbiased algorithms simply do not exist. Choices have to be made – and are made every day – by their designers.

a. Why we need to talk about the different forms of algorithmic fairness, rather than just the principle of fairness

Fairness is literally the "quality of giving to each person what is due to him or her by reference to the principles of natural justice."[23]

But,"natural justice" is neither written nor consensual. Each theory of what is just leads to its own definition of fairness. The definition of a fair decision depends on the culture, the sector of application, or the objectives one sets for oneself. And as an additional difficulty for algorithms that learn from historical data, this definition varies over time in each culture. For example, there are two main categories in the definitions of fairness: group fairness and individual fairness.

22 An algorithm is neutral when it represents reality faithfully. The algorithm is loyal when it meets the requirements set by the designer. A fair algorithm is one that makes fair decisions.
23 Definition « équité » (fairness) in the *Larousse* (translation ours).

The different forms of fairness

Fairness can be conceived in two different ways:

- **individual fairness,** which ensures that individuals with similar profiles are treated in the same way;
- **group fairness,** which ensures that the decision-making process does not arbitrarily disadvantage a certain group (the fairness generally used in algorithm testing).

The fairness that we need to concern ourselves with depends on the assumption we make about our ability to measure the world. University admission decisions are a good example of this. With or without an algorithm, the fairness of a decision requires a comparison of:

- the decision itself (whether or not to admit a high school student to a course);
- the criteria considered relevant to the decision (the student's motivation, curiosity, potential capacity to work).

Fairness can cease to exist between individuals when the decision is not based on the same criteria for everyone. But it can also cease to exist when the way the criteria are measured is to the disadvantage of certain groups.

Cover letters, for example, serve as an approximation of the motivation of applicants. This approximation is not neutral when it favors those who can seek extensive advice on how to "sell" themselves, and who are more often children of managers and teachers than of clerical employees or blue-collar workers.

If it is not possible to measure the relevant criteria directly, and if this measure is not as reliable for some groups as for others, then one may seek to correct the process through group fairness constraints.

The two types of fairness are inherently incompatible. Indeed, it is not mathematically possible to treat all individuals with identical characteristics in an identical way, while at the same time ensuring fair treatment between different groups, taking into account their divergent histories and their consequences. Faced with an unbalanced outcome in favor of one group, the simplest way to restore group fairness will involve being more stringent with respect to the individuals in the favored group.

The use of individual fairness and group fairness varies significantly across cultures and between fields of application. In the United States, affirmative action for university admissions based on ethnic background is both common and accepted practice. Fair treatment is then defined as group fairness. Conversely, in France, the use of competitive examinations and standardized tests refers to the ideal of identical requirements for all students, regardless of their origins. Both approaches reflect political and historical balances. Both are seen as conventional wisdom in their respective countries.

Moreover, these choices are not timeless and are regularly challenged. A group of students of Asian origin thus filed a complaint in 2019 against Harvard University, accusing it of favoring students of African-American and Hispanic origin in its recruitment processes, in the name of diversity. In the end, the judge found in favor of the university, stating that in order to achieve real diversity[24], it was still too early to ignore ethnic origin as one of the recruitment criteria. As Raja Chatila, Director of the Institute for Intelligent Systems and Robotics (ISIR), said in a recent article, "thinking about the ethics of autonomous systems refers to the way we see the world."[25]

It is illusory to think of defining a type of fairness that would apply in all circumstances and without taking into account the context. Recourse to legislation will remain futile given the extremely complex nature of this political hot topic. At most, one can hope for fairness in certain specific areas, such as the education or health sectors. It will ultimately be the responsibility of the developers and managers of a company or administration using an algorithm to decide which type of fairness to implement.

Faced with this diversity of definitions of fairness, there are commonly three possibilities for the algorithm: anti-classification, classification parity, and calibration logic. These three logics favor sometimes individual and sometimes group fairness.

24 Hartocollis,A. *Harvard Does Not Discriminate Against Asian-Americans in Admissions, Judge Rules*, The New-York Times, 1er Octobre 2019, mise à jour le 5 November 2019.

25 Garreau M, et Gateaud P. *Interview with Raja Chatila:Penser l'éthique des systèmes autonomes renvoie à notre façon de voir le monde*, L'Usine Nouvelle, December 2018 (translation ours).

Formal definitions of algorithmic fairness

Example of a university admissions algorithm and its fairness with respect to a protected variable: gender.

Anti-classification: The algorithm ignores the gender variable, and includes only variables not related to gender (written and/or oral results obtained for the high school diploma or results of continuous assessments, assuming of course that these variables are not affected by gender). This method pursues individual fairness between all candidates, whatever their gender. It is a case of fairness through ignoring sensitive variables.

Classification fairness: The algorithm is constrained in such a way that the proportion of false positives (students admitted who are found not to have the required level) is the same for male and female high school students. The parameters of the algorithm are adjusted in such a way that it will make the same number of errors for each group. This is another form of group fairness.

Calibration logic: The algorithm is constrained in such a way that for male and female high school students with similar grades and abilities, the admission results are completely independent of the protected variables. Unlike anti-classification, the algorithm is programmed to respect this independence between protected variables and assessment of student performance, even if this is to the detriment of some applicants.

b. Algorithmic fairness: an ideal difficult to reach

Once a definition of fairness is selected and applied, it remains complex to assert that an algorithm is fair. Indeed, it is difficult to guarantee the objectivity of the criteria used by the algorithm in its operation.

An algorithm seeks to make (or aid) decisions based on relevant criteria. But these criteria are not always directly measurable: one cannot directly measure the intelligence, potential or curiosity of a male or female high school student. We therefore have to rely on measurable characteristics, which are necessarily approximations. For example, the grades obtained in high school or the student's cover letter. Certain biases and stereotypes are inherent in such approximations.

This problem is particularly visible in the field of policing and preventive identity checks. How does one decide which people to check in an airport or at the entrance to a sports stadium or concert venue? If they are unable to objectively measure the level of dangerousness of a passer-by, the security forces will generally use criteria that are necessarily imperfect. Some might use intuition; others might focus on young men alone, or on an impression of dangerousness. Probably for fear of bias, others again will select people at random. This is particularly the case for customs controls at some airports in Mexico, where the travelers checked are partly randomly chosen.

Beyond the definition of fairness, it is therefore necessary to question the approximations made in the choice of criteria used by algorithms.

c. A fair and efficient algorithm, the squaring of the circle

A decision-making or decision support algorithm optimizes a result based on input data and an objective. The definition of this objective is the core of the algorithm.

This optimization is rarely done without imposing constraints on the algorithm, "barriers" not to be exceeded. In the case of Facebook, for example, the main objective of the advertising recommendation algorithm is to maximize the number of clicks on the ads displayed to the user. However, other constraints could also be taken into consideration: maintaining a form of diversity in the ads displayed, banning certain types of content (for example, ads for dating websites displayed for children or young teenagers), limiting political advertising, etc.

There then exists a conflict between the primary objective and the secondary constraints. This is particularly the case when fairness constraints are imposed on the algorithm: all other things being equal, these constraints result in inferior performance (with respect to the initial objective). The fairness of an algorithm is often to the detriment of its effectiveness.

Refusing to display dating site ads for young teenagers on Facebook means agreeing to reduce the number of clicks. This results in less revenue for Facebook and degraded performance of the algorithm.

The combat for a YouTube algorithm that is more ethical... and less efficient

The mission of the NGO AlgoTransparency is to inform the public on the functioning of algorithms influencing our access to information. In particular, it has actively worked on YouTube's video recommendation algorithms and has publicly questioned the negative impacts of the objectives assigned to them.

YouTube's advertising algorithm, when recommending divisive and radical content, follows the optimization logic it was given: maximize the number of recommendations that will be followed by a click from the user and a viewing of the video, which will generate advertising content.

It turns out that divisive content has a higher propensity to catch the eye and the attention of humans in general, and notably of YouTube users. The bias of the YouTube algorithm toward certain content is therefore completely voluntary.

Reducing this bias by limiting the distribution of divisive content therefore implies, if the objectives of the algorithm remain the same, a reduction in its performance.

If fairness is not taken into account upstream, then any correction will have a cost. It will require a trade-off between the performance of an algorithm and compliance with additional criteria such as ethics or fairness.

However, this trade-off is not fixed, and its criteria are not automatic. An algorithm for recommending targeted advertising for the purchase of men's razors, optimized to target only men, would mechanically have a higher false positive rate for women than for men. The recommendation algorithm is based on a gender recognition algorithm. The latter will be calibrated to recognize all women, thus avoiding showing them the advertisement, which would be an "unnecessary" expense. It will be less strict in the recognition of men, however. Not recognizing a man's gender implies not showing him the advertisement, which is a missed opportunity, but has less serious consequences with respect to budgetary constraints.

False positives for men (the algorithm thinks it is a man when it is a woman) will therefore be avoided as much as possible, while false positives for women (the algorithm thinks it is a woman when it is a man) will be more easily tolerated.

The simplest way to achieve gender equality would be to degrade the accuracy of the algorithm for men. But this would be counterproductive with its objective: maximizing the effectiveness of advertising spending to sell men's razors.

Forcing, ex post, software to respect fairness criteria necessarily leads to a reduction in its performance with respect to its main criterion, because constraints are added to its optimization function. While this may seem necessary in the field of justice, the answer is less obvious for an algorithm detecting cancer from X-rays: should its fairness be privileged over its performance in predicting cancer in certain patients?

d. Loyalty and neutrality, two complementary but imperfect approaches to fairness

Loyalty and neutrality of the algorithm are often waved as concepts allowing ethical algorithms. Just like equity, these concepts are limited and could not be sufficient to attain unbiased algorithms.
The neutrality of an algorithm consists in ensuring that it gives a faithful and identical representation of reality: the algorithm's decisions must correspond to the actual reality. If it is a question of selecting candidate résumés, then the system should propose the same proportion of male and female candidates as that existing in the database of candidates today.

A neutral algorithm could thus include, by construction, all the biases present in our society. Some will argue that this is normal, after all, because it is not the role of the algorithm to correct the problems of society. Others, however, will argue that given its ability to replicate at scale and disseminate a pre-existing bias, the algorithm has a role, and even a responsibility, in this regard.

The concept of loyalty refers to the user. It is a question for an algorithm to respect not the reality but rather the expectations of the users and consumers of the algorithm (which are different from those of its designer). This approach immediately raises the question of the identity of the users and the definition of their expectations. Is the user of facial recognition software the police officer using it to find a criminal, or the passer-by filmed and identified in a train station when taking his train? What are the expectations and how can the software developer anticipate them?

The loyalty of an algorithm depends fundamentally on the expectations of people with very different cultures. In France, one would expect an algorithm promoting job offers to ensure a balanced promotion among men and women. Another country might consider that this attitude would prevent the algorithm from choosing solely on the basis of professional criteria.

e. Use of an unbiased algorithm for recruitment

A large temping company uses an algorithm to recommend résumés to its clients looking to recruit staff. The company is sensitive to the issue of algorithmic bias. As such, it tests its algorithm to make sure that it does not discriminate according to a few basic criteria, including age and gender.[26]

The algorithm therefore measures, for each profession, the number of recommended résumés from both men and women. From there, if an imbalance is found, three possible solutions exist:

- adjust the result to achieve the gender balance historically observed for this type of job;
- adjust the result to achieve the gender balance observed in this profession at a level larger than the company (department, region, country, etc.);
- adjust the result to achieve an arbitrarily set gender balance (40-60, 50-50, 60-40, etc.).

The decision on the type of response to be adopted is a corporate decision, which is up to the management. It must be explainable, maybe not to individuals, but to a third authority. There is not necessarily a 'right answer', as each of these approaches might seem fair.

If there is an adjustment in the results, this is inevitably at the expense of the matching score of the proposed candidates and is openly accepted by the company.

In the absence of an obligation on companies, it is clear that competition will nevertheless induce them to limit the adjustment of results. However, as public opinion is becoming increasingly sensitive to these issues, it is also likely that reputational risk will form an incentive to correct obviously biased results.

26 In the United States, where the collection of sensitive data is authorized, many other parameters are tested: religion, political opinion, gender, etc.

III

MANY LAWS EXIST AGAINST DISCRIMINATION… AND SHOULD BE APPLIED BEFORE VOTING NEW TEXTS SPECIFIC TO ALGORITHMS

'Don't use a sledgehammer to kill a fly'

This English maxim should be kept in mind when dealing with the subject of algorithmic bias. In light of American scandals, it is indeed tempting to call for legislation to regulate algorithms. Nevertheless, the uncertainty about the extent and form of their future development and the low number of established cases of algorithmic bias in Europe and France call for caution.

Moreover, the strong international competition in the digital economy poses the risk of France and Europe being downgraded in the event of inadequate regulations on algorithms, with significant consequences for innovation.

Finally, substantial legislation already exists with respect to discrimination and the digital field. This provides numerous means to limit the risks of algorithmic bias, notably through transparency obligations and the possible remedies.

A. Anti-discrimination laws apply to algorithms

Discrimination existed long before the digital age and many regulatory texts in Europe and France prohibit it. Thus, French law[27] lists 25 criteria such as age or gender, the use of which is considered discriminatory in situations of access to employment, public and private services, or social protection.

27 Law 2008-496 of May 27, 2008 on diverse adaptation measures towards Community law concerning the fight against discrimination.

At the European level, no less than five European directives define the safeguards against discrimination, whether in the world of work or in access to services[28].

Where these texts prohibit the use of criteria such as gender in access to employment, this applies to recruiters, whether human or algorithm. From this perspective, the legal arsenal currently in force is transposable to algorithms and their possible biases.

The 25 discrimination criteria recognized by the French Law of May 27, 2008

The law prohibiting discrimination defines 25 criteria the use of which is prohibited in seven different situations.

The 25 criteria: In total, 16 of these criteria are recognized at European level. These are age, sex, origin, real or supposed ethnicity, health status, pregnancy, disability, genetic characteristics, sexual orientation, gender identity, political opinions, trade union activities, philosophical opinions, and religion.

Nine additional criteria are recognized under French law: family status, physical appearance, name, customs, place of residence, loss of autonomy, economic vulnerability, bank domiciliation, and the ability to express oneself in French.

The situations: Seven situations are recognized, namely access to employment, remuneration, access to public and private goods and services, access to places open to the public, access to social protection, education, and training.

Two conditions are necessary to qualify treatment as discriminatory:
- it must be based on one of the 25 criteria protected by law; and
- it must relate to one of the seven situations mentioned in the law.

28 Council Directive 2000/43/EC of June 29, 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin; Council Directive 2000/78/EC of November 27, 2000 establishing a general framework for equal treatment in employment and occupation; Directive 2002/73/EC of the European Parliament and of the Council of September 23, 2002 amending Council Directive 76/207/EEC on the implementation of the principle of equal treatment for men and women as regards access to employment, vocational training and promotion, and working conditions; Council Directive 2004/113/EC of December 13, 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services; Directive 2006/54/EC of the European Parliament and of the Council of July 5, 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation.

B. Existing digital laws limit the possibility of biases

At the European level…

Beyond the fight against discrimination, laws governing the digital field help to regulate practices. Algorithmic biases are not the subject of a legal corpus as such. Nevertheless, the laws governing the handling of personal data limit the risks for European citizens to be victims of algorithmic biases.

The European Union renewed its personal data protection framework in 2016 with the General Data Protection Regulation (GDPR)[29]. This Regulation replaces the 1995 Directive on the protection of personal data[30]. The GDPR, the first legislation in the world adapted to big data and AI, defines a number of rights and duties related to the handling of personal data. These notably include ensuring that the processing of personal data (a fortiori by means of an algorithm) is lawful, fair and transparent. The GDPR also introduces new rights for the citizen, namely the rights of objection, explanation and access.

The GDPR and new digital rights

The GDPR is a European regulation that aims to protect the personal data of European citizens. It entered into force in 2016 and introduced three fundamental rights for residents.

Right to object: Article 22 of the Regulation, entitled "Automated individual decision-making, including profiling", defines the right to object as follows:

"The data subject shall have the 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."

Any European citizen can thus object to an algorithmic decision, provided that it is fully automated and has legal consequences or significantly affects him or her.

…/…

29 Regulation 2016/679/EU of April 27, 2016.
30 Directive 95/46/EC of October 24, 1995, repealed on May 24, 2018.

Right to explanation: The right to explanation obliges the user of the algorithm employing personal data to inform the citizen subject to an algorithmic decision. This right is affirmed in Article 13 paragraph (2)(f) of the Regulation, which defines the information to be provided and in particular:

"the existence of automated decision-making, [...] and, [...] meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."

Since the entry into force of the GDPR, the explainable nature of algorithmic decisions is therefore crucial.

Right of access: The right of access defined in Article 15 allows each citizen to request the access, rectification or erasure of personal data concerning him or her. Everyone can therefore act to obtain precise information on the data used for an automated decision when the presence of algorithmic bias is suspected.

These new rights provide the citizen with some means of recourse against algorithmic decisions that are allegedly biased. However, they require a proactive attitude on the part of the citizen to obtain more information and to challenge a biased algorithm. In the digital field, obtaining this proactiveness can be difficult. One example of the reluctance to take an interest in these tedious and complex subjects is the lack of reading by Internet users of the conditions of use of personal data that are now displayed before entering a website.

While the GDPR does lay the groundwork for legislation on algorithmic bias, it nevertheless suffers from several limitations.

Firstly, it remains a fairly general text, with the possibility of multiple exemptions. The case law of national courts and the European Court of Justice will define its real scope. Thus, national legislation can limit these new rights in the fields of national security, national defense, public security, justice, and for general interest objectives in the monetary, budgetary, fiscal, health and social security fields[31]. There thus exists a specific directive on the protection of personal data for police and judicial activities[32].

31 Article 23 on the limitations of the Regulation.
32 Directive 2016/680 of 27 April 2016.

Secondly, the GDPR does not specify anything about algorithms reproducing societal biases on a large scale, which would probably not be contrary to the Regulation in terms of personal data.

In addition, by limiting the collection and retention of personal and sensitive data, the GDPR restricts the ability to test for the absence of algorithmic bias.

Finally, European law only targets algorithms when they use personal data, or make decisions about individuals. However, some algorithms can have impacts on citizens without using personal data or making decisions that directly affect them.

When the CNIL warns against algorithms that do not use personal data

The CNIL was tasked with formulating its opinion on the public debate concerning the ethics of digital technologies. In its summary document published in 2017, it warned of the legal void concerning algorithms that do not use personal data. Despite the absence of any direct decision affecting citizens, these may represent risks.

Let us take, for example, an algorithm defining the types of meals consumed in the school canteens of a given region based on non-personal data, such as the history of the menus served by the canteens. This algorithm could have biases, such as a bias favoring certain types of food beyond what is desirable, or a bias attributing certain meals more regularly to one type of high school than to another, without any explicable reason. It is not certain today that this bias, collective in nature, would fall within the scope of personal data legislation, despite its impact on individuals.

Beyond this warning, the CNIL has brought to light the principles of loyalty and vigilance of algorithms, in both the private and public sectors. While the first is now referred to in several legal texts, the second, more recent, consists in organizing, by technical and human means, the regular questioning of the operation and results of an algorithm, predictive or otherwise.

This recommendation sheds light on the priorities that the French ecosystem wishes to set for itself in terms of algorithms: loyalty, for the person concerned but also with respect to society, and constant vigilance before and especially during the use of algorithms.

Despite its limitations, the GDPR is a tool that wields powerful levers to ensure its implementation. It is by nature extraterritorial in scope. Indeed, it applies to any entity or person handling the personal data of European citizens. A US digital company with users who are European citizens is thus fully concerned by the provisions of the GDPR. Penalties for non-compliance with its provisions are substantial and can amount to €20 million or 4% of worldwide turnover. The GDPR thus boasts substantial means to encourage actors and stakeholders to respect the new digital rights of citizens.

… and at the national level

Beyond this European approach, with all essential aspects being defined in the GDPR, French national law in the digital field also provides many answers on the subject of algorithmic bias.

The French Digital Republic Bill of 2016[33] authorized automated decision-making in the public domain[34] under two conditions: citizens must be informed that the decisions concerning them have been automatically taken and can, at their request, obtain information on the algorithm used. This notably encompasses the degree and method of contribution of the algorithmic processing to the decision-making, the data processed and their sources, the processing parameters and, where appropriate, their weighting, applied to the situation of the person concerned and the operations carried out by this processing.

This law has a fundamental impact on automated individual decision-making: it is considered illegal in cases where the algorithm is not technically explainable or where certain elements of its operation cannot be communicated for legal reasons (tax secrecy or defense-related matters, for example). Full transparency of the algorithm is therefore mandatory once the algorithm takes an individual decision. This is beyond the scope of the European regulation.

33 Law 2016-1321 of October 7, 2016 (link), the role of which was notably to update the 1978 Data Protection Act, the founding law on the protection of rights in the digital field, and which includes certain specificities in terms of algorithmic bias.

34 This type of algorithm already exists, for calculating the amount of taxes due, for example.

When the Constitutional Council censors machine learning

The Digital Republic Bill makes it unlikely that machine learning algorithms will emerge for individual public decision-making. It would notably be complicated to guarantee that the algorithms could be fully explained for these technologies with decision trees that are in essence difficult to interpret.

Nevertheless, the French Constitutional Council has further restricted the possibilities of using machine learning for public decisions.

In its opinion delivered on June 12, 2018 on this same law, it specified that a self-learning algorithm, i.e. one that revises its own operating rules, without the opinion and supervision of the data controller, cannot lead to automated individual decision-making.

In other words, self-learning algorithms are not constitutionally authorized to make decisions on behalf of the public authorities. The risk of bias in this area is therefore more limited.

While the Digital Republic Bill reinforces the transparency requirements for public algorithms and restricts the use of automated decision-making, it does not add constraint to existing frameworks for privately-owned algorithms, which are subject to the GDPR.

Finally, and through sectoral legislation, France is gradually adopting practices to be observed in the use of algorithms. This applies in particular to the health care, transport, and information sectors.

Article 11 of the draft law on bioethics, tabled in the National Assembly on July 28, 2019[35], introduces obligations when using algorithms for the processing of mass data for preventive, diagnostic or therapeutic purposes. In particular, the health professional must inform the patient of the use of an algorithm and its operating procedures. Moreover, the health professional will retain the possibility of modifying the parameters of the algorithm, the behavior of which will be recorded.

35 French Bill no. 2187 relating to bioethics, July 24, 2019.

The Mobility Orientation Law of December 24, 2019[36] sets out transparency obligations for platforms, particularly with regard to the criteria used by the algorithms leading to decisions on the remuneration of drivers or the location of the missions assigned to them.

Finally, the proposed law to combat hate on the Internet[37] should also tighten the rules on algorithms for fake news. Following a logic of due care, the text imposes on platforms to remove all hateful content within 24 hours of publication. The ex-post verification of possible errors in a content sorting algorithm is a provision which has also been adopted in other countries such as Germany[38], and which makes it possible to ensure compliance with the law without having to explain or fully understand the operation of the algorithm.

However, these different laws are limited in terms of algorithmic bias: they deal more with measures that come after the use of a biased algorithm, in particular by establishing rights of appeal and transparency. They do not propose anything ex-ante, i.e. before a possibly biased algorithm goes into production.

C. The developing specificities of European and French law in relation to the United States in the field of algorithms

Numerous protection measures exist against algorithmic bias, whether they concern the fight against discrimination or the laws governing the digital field (GDPR, Digital Republic Bill, etc.).

Despite the few examples of algorithmic bias in France and Europe, the laws differ from Anglo-Saxon models on several points.

Firstly, the use of data concerning the ethnicity of individuals is an important point of divergence. Contrary to the United Kingdom or the United States, the collection and use of such data are considered unconstitutional in France. While the 1978 Data Protection Act[39] includes exceptions to statistical collection in order to authorize the

36 French Bill and separate report of the Mobility Orientation Law no. 2019-1428 of December 24, 2019.
37 French Bill no. 1785 of March 20, 2019 on combating hate on the Internet.
38 Orsini,A. *Discours haineux : les réseaux sociaux risquent 50 millions d'euros d'amende en Allemagne* dans Numerama, January 2, 2018.
39 Exceptions to this prohibition exist in the medical and security fields, or for data that will be anonymized.

collection of data on health, political or religious opinion, or again sexual orientation, the collection of data relating to ethnic origin is specifically excluded. Such collection is, however, the cornerstone of the fight against bias and of affirmative action in the United States.

The French Universalist tradition refuses to identify ethnic communities within the Republic. In 2007, the law on immigration control, integration and asylum[40] contained a provision authorizing the collection of such data. That provision was then censured by the Constitutional Council[41], invoking Article 1 of the Constitution, which lays down "the equality of all citizens before the law, without distinction of origin, race or religion." In the commentary on its decision, the Constitutional Council considered that studies can be conducted on either objective data (place of birth, name or nationality, etc.), or subjective data (the sense of belonging, etc.). Assigning individuals an ethnic identity is, however, formally prohibited.

Secondly, the USA does not yet have any federal legislation concerning algorithms or the protection of personal data. Although the Federal Trade Commission does issue some recommendations based on sectoral laws (financial markets, youth protection, etc.), it is the individual States that are responsible for dealing with these subjects.

Illinois recently passed a law requiring companies using decision support algorithms for recruitment purposes to inform applicants of the use of an algorithm, to explain how it works and the criteria analyzed, and to seek the consent of the applicants.

Massachusetts and California have also implemented laws similar to the GDPR with the introduction of transparency obligations and the citizen's right of access. These texts are nevertheless limited to the processing of personal data and do not concern any automated decisions.

At the federal level, the idea of a specific regulation on personal data is not consensual. The Internet is considered – in particular by the Supreme Court – as a public space. As such, the information published on the Internet is considered by many people in charge to be public. The use of personal data and algorithms is thus more closely regulated in texts on consumer law or anti-discrimination law.

40 French Bill n°2007-1631 of November 20, 2007.
41 Decision of the Constitutional Council, 2007-557 DC.

The US Algorithm Accountability Act

While the claim to privacy on the Internet is disputed, discrimination caused by algorithmic bias is a frequent subject of controversy in the United States. The scandals involving Facebook's real estate ads and Amazon's résumé rating algorithm prompted the Democrats to propose an Algorithm Accountability Act to the Senate in April 2019.

If this text were to be voted as it stands, it would oblige the federal administration, the States, and companies with a turnover of more than 50 million dollars and containing information on more than one million people, to carry out impact studies on their decision-making or high-risk automated decision support algorithms.

High-risk algorithms are those affecting sensitive elements of privacy such as work performance, health, personal life, location, or using data such as race, gender, religious or political opinions. This also includes the surveillance of important public spaces, and thus potential facial recognition uses.

These impact studies should verify the behavior of the algorithm with the following criteria: accuracy, fairness, bias, discrimination, privacy and security. In the event of negligence, the entity would have to take remedial action and the FTC would be given the power to impose sanctions.

This law could gain support from the Republicans who are displeased with Facebook, Google and Twitter, accused of political bias in favor of the Democrats in their recommendation algorithms.

D. The application of existing law to algorithms is today imperfect and difficult

Both discrimination and the digital industry are already substantially covered by the law. The entities in charge of its application include both bodies specialized in these issues, such as the French Ombudsman or the CNIL, and sector regulators such as the ARCEP for the telecommunications industry, the CSA for the media, or the ACPR for the banking and insurance sectors. In both cases, it is currently difficult for these entities to fully enforce existing law.

For sector regulators, conflicts of interest can limit the ability to combat algorithmic biases. Indeed, these entities often have a main objective that is far removed from this type of subject.

For example, the insurance industry has a dedicated regulator, the ACPR, which looks not only at the variables used in risk assessment and pricing models and algorithms, but also at their quality, as financial stability depends on the proper assessment of risks. Rather than unbiased algorithms, the ACPR may thus prefer biased but accurate algorithms in order to ensure financial stability. The conflict between these two objectives therefore probably requires to think about a sharing of responsibilities concerning the regulation of biases in insurance algorithms.

As regards crosscutting regulators such as the French Ombudsman and the CNIL, the difficulties today reside in their ability to effectively regulate a colossal number of entities. It is difficult for the CNIL, with a budget of 18 million euros and 210 staff members, to fully audit the 80,000 bodies that must today appoint personal data officers?[42] In 2018, only 310 such audits were carried out.

The CNIL is today facing real difficulties in terms of resources to implement the GDPR. Adding new prerogatives in terms of algorithmic bias via new regulations would only aggravate this situation.

This is why it is essential today for France to focus on the full implementation of existing regulations on discrimination in the digital field. Indeed, there are many provisions that make it possible to prevent the negative effects of bias, provided that they are gradually applied to new algorithms.

42 French Senate, Projet de loi de finances pour 2019 : Direction de l'action du gouvernement, publication officielle et information administrative.

IV

RECOMMENDATIONS

For a long time now, computer scientists and statisticians have known about biases, because technical biases are inherent to algorithms. However, managers, civil servants, judges and citizens that are now facing them are discovering this issue, often through American scandals. Algorithmic biases are a little-known phenomenon in France, the most emblematic and talked-about cases being American. Within the computer science community, this subject is not new since technical biases are inherent to any algorithm. However, managers, civil servants, judges and the general public are now discovering this problem, sometimes with horror. There is today a shared observation on the increasing presence of algorithms in our lives and on their possible biases. Two conflicting approaches feed the public debate on this subject.

Some argue that algorithms must adapt to our society, and if they are not able to prove that they are unbiased, they should not be used. The right tool would then be a law on algorithms, which would prohibit or restrict their use in many cases. The corollary of this would be a strong regulator, whose mission would be to verify the lack of bias in the algorithms in use. Proponents of this approach no longer hesitate to label algorithms as mathematical weapons of mass destruction[43]. For them, the benefits of algorithms are not worth the risks.

The other side argues that it is, on the contrary, more necessary than ever to take advantage of the innovations made possible by digital technology. Even if algorithms contain biases, they will always be less biased than humans. Moreover, if we do not start developing these technologies, the United States and China will do it for us. In addition to being biased, these technologies will then also be controlled by entities outside Europe, as is already the case for all algorithmic services produced by GAFA. The digital gap would be impossible to fill and its economic impact would be fatal to European prosperity. It would therefore be best not to propose new regulations and to let public and private actors develop algorithms as they see fit, thereby facilitating innovation to the greatest extent possible. After all, the benefits far outweigh a few biases.

Both approaches are, in several respects, excessive. While it is clear that algorithmic biases present a danger, they must be set against the opportunities that the digital

43 O'Neil C., *Weapons of Math Destruction*, Penguin Books, June 2017.

revolution offers. Algorithms are first and foremost an excellent way to discover and reduce existing biases in humans. They force us to formalize the choices we make in terms of fairness. Is algorithmic bias harmful if it is much less so than an existing bias? Algorithms represent the large-scale deployment of locally observed behaviors. No-one will accept to see the massive spread of biases in our country resulting from biased learning databases. The examples from the American justice system are extremely alarming, and rightly so.

The challenge is therefore to strike the right balance between supporting innovation and ensuring the adequate framework for managing the risks of discrimination. We are convinced that a step-by-step approach is necessary.

The use of algorithms and our knowledge of their biases changes all the time. We must therefore develop our capacity to understand, detect and react to these biases – with one objective, namely to boost confidence in the operation and fairness of algorithms in order to accelerate their dissemination.

Our recommendations are threefold: prevent the introduction of algorithms biases, notably through training; detect their presence via testing within the organizations; and have high-risk algorithms evaluated by third parties.

A. Proposals left on the side

Algorithms and their defaults are a particularly worrisome subject for many Europeans, particularly in France. Algorithm biases are perceived as a first step towards a society in which equality would have disappeared because of technology – optimization, statistics and technological progress. The loss of control towards programmers is the first thing coming to European minds when asked about algorithms, according to a survey[44] from the Bertelsmann foundation in 2018, with the French being the most worried. They are about 72% to think that a recruitment process assisted by an algorithm would primarily be a threat, according to a survey[45] cited by the CNIL in 2017. Algorithmic biases are rightly seen as the beginnings of a society where technology would have sacrificed equal opportunities on the altar of optimization, statistics and technological progress.

44 Grzymek, V. et Puntschuh, M. *What Europe Knows and Thinks About Algorithms*, Bertelsmann Stiftung, February 2019.

45 Cited in *Comment permettre à l'homme de garder la main?* , summary of the public debate tasked to the CNIL by the law for a Digital Republic, December 2017.

Faced with this situation, there are many calls to establish strong, quick and emblematic decisions. However, we believe that such an important subject should not give rise to hasty initiatives. There are therefore a couple of proposals that will not be made in this report.

• Non-proposal #1: A law on algorithmic bias

The ambition of the new European Commission to propose initiatives to promote the emergence of ethical, responsible and unbiased AI seems to us a very positive step in the right direction. Moreover, we fully agree with the approach of concentrating efforts on potential high-impact algorithms. Nevertheless, we believe that it is premature at this stage to work on a directive governing AI and the ethics of algorithms[46] with respect to algorithmic bias.

As explained in the first part of this report, established cases of algorithmic bias in Europe and France are still few and far between. It is therefore difficult to write balanced regulations on problems for which there is no hindsight. Legislating without having taken the time to observe and analyze the phenomenon in detail means taking the risk of over-regulation that is harmful to the digitization of the economy and to innovation. The alternative, under-regulation, could let liberty and equality violations occur, leading to sudden legislation initiatives following scandals.

The second major obstacle to a law on algorithmic bias lies in the extraordinary diversity in the uses of algorithms. The banking, insurance, automotive, health care, recruitment, advertising, justice or law enforcement sectors are some of the areas where a massive deployment of algorithms will be necessary. As such, how would it be possible to define rules that apply to everyone, without taking into account the specificities of each sector? The definition of bias and what constitutes fairness will inevitably be different depending on whether we are talking about an algorithm for automated driving, the development of a chemotherapy protocol, or targeted advertising.

Promoting a "law to prevent and combat algorithmic bias" in 2020 would, in our view, constitute an inappropriate approach. The mobilization of civil society actors and companies seems much more favorable to us at this stage while we wait to gain a clearer vision of the situation. This approach in no way alters the State's capacity to take regulatory initiatives for the most critical use cases, which are, of course, numerous.

46 European Parliament, legislation train Schedule, Communication on artificial intelligence for Europe, May 2018.

• Non-proposal #2: State control of algorithms

The GDPR and its implementation in France by the CNIL have shown the limits of a regulation model. It is an illusion to expect the CNIL to be able to check and authorize all of the processing of personal data that takes place in French society. The obligation for the data protection officer to carry out an impact assessment instead of obtaining a statement by the CNIL is the consequence of a brutal reality: the digital world is far too vast and changing to be controlled ex ante by an independent authority, even if the latter were to be endowed with considerable resources.

Believing that the difficulties met when letting a regulator operate ex ante control over the digital sphere only concern personal data and that the situation would be different for algorithms is just as illusory. This is why we consider that expecting the State to check the absence of bias in algorithms is neither feasible nor desirable. It should now be considered that this role of controlling algorithmic biases should be at least partially transferred to companies implementing algorithms and to research and certification laboratories, or to third parties along the model of firms that audit corporate financial accounts, for example. This decentralized control should be accompanied by greater accountability from actors regarding the consequences of algorithm bias.

B. Prevent bias by implementing best practices and increasing training for those that create and use algorithms

In order for AI technologies to be successfully integrated in and beneficial to our society, we recommend that the entire chain of actors involved in the production of or affected by algorithmic decisions be properly trained in the risks of bias and discrimination, and understand the risks and benefits of deploying algorithms. The deployment of best practices in companies and administrations, particularly in terms of diversity within teams, is also essential.

Figure: The algorithm "production chain" (large arrow with the different actors and the type of bias that can be introduced at each stage).

Some algorithms are developed internally within the same organization that then deploys them (e.g. Amazon's recruitment algorithm). Others are purchased from suppliers and deployed by organizations that did not develop them (e.g. the PredPol crime prediction algorithm used by the Los Angeles Police Department).

An entire production chain for algorithms thus exists:

Management: Technicians are supervised by managers, who make choices about the resources to be allocated to the project and the objectives to be optimized. These managers can decide that the detection of potential biases is not a priority given the estimated risks, and can set a strategic objective for the algorithm that reproduces societal biases.

Development: The computer developers or data scientists who code the algorithms are their architects. Developers don't usually works in isolation: in the spirit of software development, machine learning is widely published in open source on the Internet. Large companies develop their own tools, but then make them available. In 2015, Google published the TensorFlow software library that allows everyone to create their own neural network.
When in 2018 a study conducted by the MIT Media Lab proved that IBM and Microsoft algorithms recognized the gender of 99% of light-skinned men but only 65% of dark-skinned women, the New York Times questioned the level of diversity within the teams that had designed these algorithms.

Data: Learning data, crucial even before using the algorithm on one's own data, are rarely collected internally or in isolation. Databases are published by researchers. ImageNet, the database developed at Princeton University, provides access to some 14 million hand-annotated images across 20,000 categories, and allows them to be used to train visual recognition algorithms.
Nevertheless, any bias in the database can create biases in numerous algorithms. AI Now's ImageNet Roulette project, for example, revealed that learning on ImageNet could result in multiple biases (black people with selfies labeled "criminal" or "convict"). ImageNet removed over 600,000 images from its database in an attempt to resolve this problem.

.../...

Data Labeling: In addition, there is a whole data labeling industry that covers everything from image recognition to content moderation. The decisions of these click workers may not always correspond with user expectations, especially when a large part of the industry is relocated to low-wage countries with different cultural references. This is the difficulty of the task required of Facebook moderators, who create learning databases for moderation algorithms, and include their own biases in them.

Use: The algorithm is often only a decision aid. Judges rule on the release of a prisoner. Amazon recruiters publish a job offer. Algorithmic biases have impact only when they are not filtered or corrected by the decision-makers. Rather than imposing fairness requirements in their algorithms, some banks decide, for example, to entrust the prediction of credit risk to a person who will be responsible for correcting the algorithm's recommendation based on his or her own experience.

Feedback ("reinforcement"): Algorithms are inserted into systems that collect usage data, allowing for fine-tuning their operation and optimizing their performance. Someone who communicates with a chatbot orients the system. For example, Microsoft's chatbot Tay, launched in 2016 on Twitter, quickly began to post inflammatory and offensive tweets, including racial slurs. It simply reproduced the behavior of the Internet users who were "testing" the system by insulting it. Its Twitter account was closed in under 24 hours. Since Web 2.0, where interaction is the rule, any feedback loop on smart systems is likely to create a bias in the algorithm.

With so many players, the question of liability in the event of algorithmic bias is crucial. If predictive policing is racially biased, who is responsible? Is it the user (the LAPD officer), the manager/buyer (the LAPD Chief Information Officer), the developer, or the provider of biased data?

- **Proposal #1: Deploy good practices to prevent the spread of algorithmic bias (internal charters, diversity within teams)**

Algorithmic biases present a real danger for citizens but also for the companies and administrations that design or deploy them. Preventing biases is far less costly than correcting them because, beyond the legal risk, the reputational stakes are considerable. It is therefore essential that each link in the chain implements best practices to prevent, detect and warn of possible biases.

Without much publicity on the subject, several French companies are gradually putting in place practices and charters to deal with the risk of bias. We recommend that all actors wanting to use algorithms follow their example by implementing a charter for the development and deployment of fair algorithms (this would of course not apply to the most basic algorithms presenting no risk, such as the ones establishing the optimal volume when listening to music in a car).

Without covering all aspects, certain points deserve to be included in these internal charters:

- the methodology requirements to ensure the quality of the algorithms;
- the properties that the developed algorithms must have (in particular so they can be later audited during deployment);
- the internal mechanisms to manage tensions between different objectives, to define fairness requirements for the algorithms and to specify their formalization.
- the internal analyses and assessments to be performed on the algorithm;

Box: Some examples of good practices

Masking sensitive variables: Some companies, particularly in the banking, insurance and recruitment sectors, isolate the protected variables of their clients (age, gender, address, etc.). These variables are no longer accessible to risk assessment or pricing algorithms. This practice is intended to reduce the risk of an algorithm using these elements as a discriminating factor. This "fairness by unawareness" approach has its limits because machine learning algorithms can easily infer the protected variables (e.g. gender via car model and color).

Comparing false-positive rates: In recruitment, large agencies systematically check that the false-positive rates in their algorithms, i.e. classification errors, are the same for different subgroups of the population. The aim is to avoid scenarios where software would be much more effective for some people than for others, even if it means lowering performance levels. This aims to avoid cases similar to facial recognition in the United States.

.../...

Monitoring algorithms after deployment: More and more algorithms are learning and evolving as they are used. Some companies therefore check that these algorithms are not introducing new biases by repeatedly testing their fairness at regular intervals.

Involving the management: Several companies stipulate very precisely the cases in which the management or the risk committee must be consulted (introduction of a new variable, trade-off to be made between performance and fairness, definition of a fairness criterion to be evaluated).

This excerpt of best practices aims to provide companies deploying algorithms with some ideas to help them implement charters for the development of fair algorithms and thus combat the risk of bias. This step is essential to change behaviors and prevent any discrimination scandals that could emerge.

Algorithms are not only deployed in companies that possess expert knowledge in this field. Some organizations purchase mature technologies to increase their productivity, to ensure interaction with clients, or for human resource purposes. However, the managers in these organizations generally have a much more limited knowledge of the risks and limitations of algorithms.

As with any sensitive purchase, buyers must have a seasoned mastery of the subject to be able to question their suppliers. Good practices will spread all the more quickly where buyers implement stringent requirements.

In terms of good practices, it is necessary to emphasize the importance of building socially and professionally diverse teams in algorithm development projects. The managers and technicians of data-centric companies have a very good understanding of the opportunities represented by the algorithms they develop for external and internal requirements. But their understanding of fairness challenges can sometimes suffer from a lack of diversity in their ranks: women are more likely than men to detect the potential discrimination risks from an online technology, while younger and older people relate very differently to online tools.

Composing a team of developers that is truly diverse in terms of gender, social background, ethnicity, or other criteria, is bound to come up against the lack of diversity in training in computer science, statistics, or data science.

- **Proposal #2: Train technicians and engineers in the risks of bias, and improve citizen awareness of the risks and opportunities of AI**

While developers seem to be on the front line when dealing with biases, it is in fact all the actors in the life of an algorithm who are concerned. These actors include scientists and business managers, but also citizens, employees, political leaders, and so on. We recommend that each of these groups benefit from training on the subject of bias and artificial intelligence. While the training of scientists and developers could focus on algorithmic bias and in particular societal biases, the content intended for the general public should offer everyone the possibility to understand the challenges inherent in artificial intelligence.

In France, most developers have been trained in applied mathematics, statistics and computer science, without having received any specialized training in the social sciences. They are taught to understand the technical challenges of designing and optimizing algorithms, not the societal challenges.

In technical terminology, bias is anything that causes the algorithm to deviate from producing the optimal result of what it was designed to do. It is limited to the biases that we call "technical", and does not include the impact that the optimal result produces on the groups of people whose data were used to "train" the algorithm, or on those whose lives will be affected by the algorithm's prediction.

Data experts require specialized training in understanding societal biases, as well as in the different notions of fairness. They will also have to learn how to adapt these general frameworks to specific situations and algorithms.

These experts will need to understand the importance of collecting a sample of learning data correctly reflecting the population that will be affected by the algorithm, beyond mere statistical representativeness that might not take societal biases into account. The problem is often that the learning data represent populations that are familiar to the data scientists who designed the algorithms. These algorithms, learning from limited examples, cannot then produce appropriate decisions for populations that are excluded from the learning sample.

Google's early facial recognition algorithms, for example, classified some black people as gorillas! The algorithm had learned using a limited set of photos that reflected the social group of the designers – white males. The database did not contain enough faces of people of color to allow the algorithm to correctly identify these people in the new images it was presented with.

Training data specialists and engineers to assemble a learning sample that correctly represents the population is one of the most important prerequisites for increasing fairness in algorithmic decision-making.

This requires above all the recourse to continuing education using the tools of the 21st century (discussions on development platforms, MOOCs), and not just university education. The latter is important, but not sufficient, notably because it is also about training people already employed.

Beyond scientists, business managers making decisions to implement algorithms must also be trained, but on different issues. They need to keep in mind priorities, such as having a diverse set of profiles in development teams and implementing procedures to evaluate the risk of algorithmic bias. In this respect, the good practices mentioned in the first proposal of this report and based on real situations that we encountered should serve as pedagogical examples in order to deploy this type of initiative more widely.
Finally, citizens, employees and political leaders must be aware of the issues involved in the development of algorithms and AI in order to be able to respond to fears and exercise a duty of vigilance.

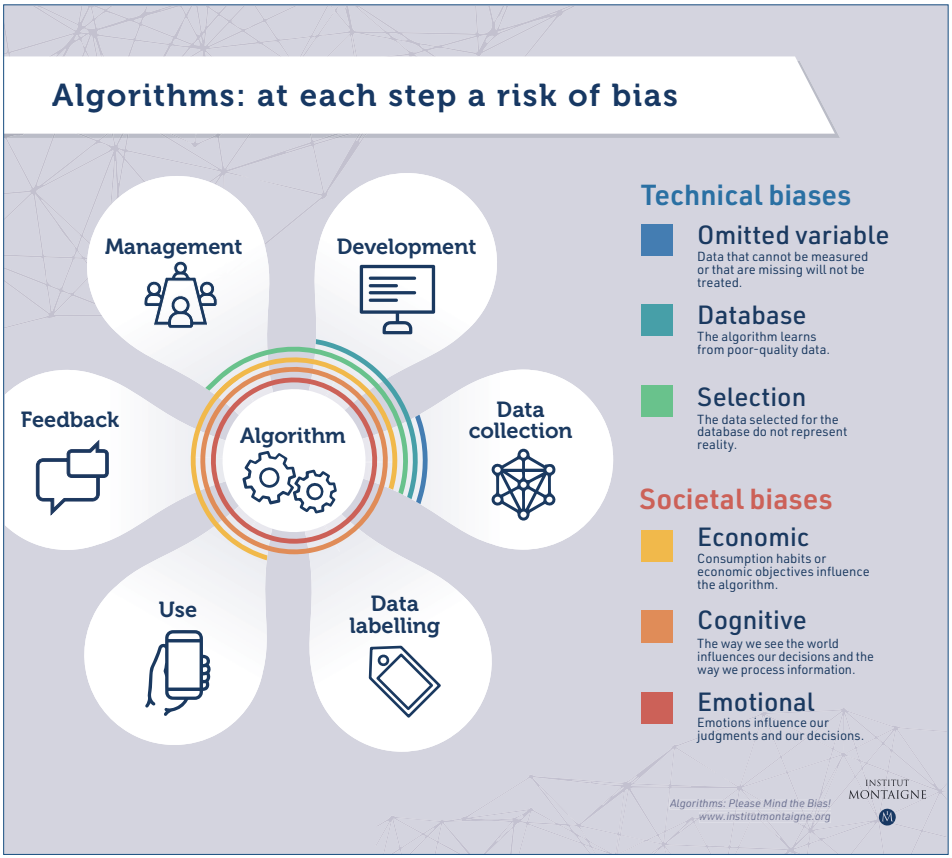
Indeed, the adoption of AI will not be possible in France without first addressing the fears it raises, which at times are close to science fiction, where machines would take control over humans. These scenarios coexist with serious concerns about the impact of AI technology on the labor market and about the way private data are used on the Internet and social networks. It is important to separate myth from reality – the general public wants and has the right to understand how algorithms are designed, what data are used, how biases will be handled and how this technology will impact their lives. Everyone must be given the means to achieve this understanding, which will then create confidence and also ensure informed vigilance where this technology is concerned.

Basic knowledge about machine algorithms should be included in school curriculums, but also be accessible to adults through continuing education. In Finland, 3.5% of the population has already received training in AI thanks to a course designed locally and specifically for them.

In France, the Institut Montaigne and OpenClassrooms, in partnership with Fondation Abeona, have developed “Objective AI”, an online training course that will be available in the first quarter of 2020. Through a series of videos, interactive exercises and examples, the course will address several fundamental topics, both technical and societal, to enable everyone to comprehend the opportunities and challenges of AI and develop critical thinking skills that will help them understand and navigate this new

technological landscape. The course will be available free of charge online, but will also be distributed by many companies to their employees.
We recommend the development of this type of initiatives aimed at citizens and public decision-makers to improve their knowledge of these subjects.

Another ongoing initiative is entitled “1 scientist - 1 class: No problem!” The Ministry of National Education has signed an agreement with Inria, the National Institute for Research in Computer Science and Automation, to develop the pilot phase of this project. This undertaking is part of the new “Digital Science and Technology” teaching program that enables tenth grade students to meet male and female researchers in the digital sciences. The students come out with a better understanding of a world transformed by digital technology, thereby nurturing interest and encouraging vocations, especially among young women. This type of initiative, the development of which we strongly recommend, should integrate the theme of AI algorithms and their impact on society.



C. Give each organization the means to detect and fight algorithmic bias

- **Proposal #3: Require testing algorithms before use, along the same lines as the clinical studies of drugs**

There have been numerous calls recently to promote explainable artificial intelligence. This is notably the case of the Villani Report in 2018, which defined the French strategy in this field[47].

However, we believe that it is difficult to simply explain all algorithms. We rather defend the need to test algorithms, to ensure that they do what they were developed for on the one hand, and to detect the presence of bias on the other. These tests would occur as in drug clinical studies, when the pharmaceutical firm verifies the harmlessness and efficiency of the drug on a sample of patients.

This approach is more effective in several respects than just being able to explain algorithms. Indeed, there are many limitations to the explainable nature of algorithms, and it is not certain that fully explainable AI is possible. Moreover, the complexity of algorithms is increasing day by day. Even an algorithm that is technically explainable can remain abstruse for the majority of people.

The limits of explainable AI

The explainable nature of algorithms suffers from many limitations. Although desirable, it is technically difficult to obtain as it is contrary to the very principle of machine learning. AI looks for data patterns automatically, without allowing the user to make sense of the sequence of calculated correlations. Worse, with current techniques, there is a real opposition between the performance of an algorithm and being able to explain how it works, reducing the prospects of this ability to systematically explain algorithms.

.../...

47 Parliamentary mission from September 8, 2017 to March 8, 2018, Report by Cédric Villani, «*Donner un sens à l'intelligence artificielle*» ("Giving Meaning to Artificial Intelligence").

Moreover, explaining an algorithm does not really meet the requirements in terms of bias. An algorithm comprising 50,000 rules, all explainable, will remain incomprehensible to the average person. It is also difficult to mobilize users on these subjects. They may not be interested in the legal terms detailing how an algorithm works, and may validate them without reading – as so many of us do. They will not seek to understand the algorithm, but rather to be reassured that it works fairly. Who wants to know how an airplane works, for example? All the reassurance we need is to know that it has passed the safety tests.

Finally, being able to explain an algorithm risks violating the concept of business secrecy, which protects to some extent the content of the algorithms (variables, weightings, etc.), which are increasingly strategic.

Without being able to explain them, it is nevertheless possible to ensure that algorithms have certain properties. Tests are much more likely to reassure us of what really matters, namely that individual were treated fairly.

These tests should be performed as much as possible by the organizations deploying the algorithms, whether private or public. These entities are in the best position in terms of the skills and means necessary to carry out these tests. Even though the tests could represent a cost and significant technical challenges, the entities concerned have a vested interest in limiting all legal and reputational risks.

Conducting such tests requires defining what the company or administration using the algorithm considers to be fair. Except in rare cases, this framework will not be set in stone by the State, as it depends on the definition and the circumstances (sector of application, criticality of the algorithm, period, etc.). Certain businesses consider that the rates of false positives on groups should be identical to within a 5% margin. In recruiting, some companies favor strict parity or give themselves room for maneuver (no less than 40% women). Others favor a parity in line with the sector, with the composition of graduates. In order to help determine these thresholds and the appropriate testing methodologies, it would be helpful to be able to count on the support of the CNIL both upstream of and during the testing process, as is the case for the protection of personal data.

Some biases are voluntary, acceptable, and the result of commercial strategies; others are not. It is therefore ultimately up to the company to position itself on what it considers to be the right definition of a fair algorithm. It then bears the legal and reputational risk, but refusing to confront these issues would only increase this risk.

- **Proposal #4: Adopt an active fairness approach – authorize the use of sensitive variables for the strict purpose of measuring biases and evaluating algorithms**

In contrast to Germany, where the unity of the German nation eventually pushed the German States to come together, it is said that in France it is the State that built the nation, ignoring and erasing differences of dialect, religion and beliefs, with more or less success. Within this movement, France has sometimes preferred to promote a certain universalism, masking its diversity in an attempt to treat everyone equally. With only some rare exceptions, ethnic statistics are prohibited and data on the 25 criteria of discrimination are extremely difficult to collect. But how can we test for bias on these 25 criteria if this data is not available?

Today, this provision poses difficulties for measuring discrimination. Sociologist François Héran, a professor at the Collège, France's most prestigious public research establishment, writes in the preface to the Trajectoires et Origines statistical survey[48], "Will we still believe [in ten years' time] that some of these questions on origins or appearances were suspected of wanting to 'undermine the foundations of the Republic,' when in reality they were modestly aimed at comprehending as best as possible the mechanism of discrimination which compromises the principle of equality?"

Excluding sensitive variables from an algorithm is insufficient. In the field of advertising algorithms, it is not necessary to know the gender of someone who buys men's shirts or women's swimsuits as this information can be guessed with only a small margin of error. Fairness by unawareness is not sufficient because algorithms can discriminate on criteria to which they do not have direct access. As explained by Stanford researchers S. Corbett-Davies and S. Goel in a paper they published on algorithmic fairness[49], the exclusion of sensitive or anti-classification variables does not effectively meet the goals of fair algorithms. In many cases, a calibration approach, i.e. an approach by which one ensures that the results obtained are indeed independent of the protected variables, is preferable.

We recommend abandoning the fairness by unawareness approach and adopting an active fairness strategy. Practically, this means measuring discrimination through bias testing by collecting data on the 25 protected criteria.

48 Survey carried out in 2009 and collecting data notably on geographical origin, nationalities prior to French nationality, and the sense of belonging in the French population (translation ours).
49 Corbett-Davies S., Goel S., *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, Stanford University, July 2018.

The algorithm is fair, when the independence of the result with respect to protected variables is guaranteed, not when the sensitive variables are excluded. It is through this approach that we will collectively be able to identify and reduce biases in algorithms and in society in general.

Nevertheless, this approach must be strictly regulated. Today, exceptions, authorizing the collection of sensitive data already exist, notably within the framework of statistical uses. We propose to extend these exceptions to the very specific case of tests to identify possible discrimination, by authorizing a reduced collection – on a sample of data – of these sensitive variables.

Where this is justified by the nature of the algorithm, the developer can then verify that the results are independent of the protected variables. The collection of such data would of course require the consent of the user, who could agree to share such information in order to help build a test base that would allow for verifying the absence of bias. The collection of such data would also require the prior completion of an impact study and its transmission to the CNIL.

As a precautionary measure, the collection would be limited to a small sample of users, to be defined. This could concern either a fraction of users (20%, for example), or a sufficiently large sample to guarantee statistical representativeness (usually a few thousand users). This limit would make it possible to avoid generalized abuse of the data while guaranteeing that the data collected is sufficiently representative. In some cases, it will remain difficult to gain the confidence of individuals and convince them that the data will only be used for testing purposes. In the case of recruitment, for example, a candidate might be suspicious of a recruiter who would want to collect protected variables, even with the assurance that these would be used for testing purposes only. Data collection could therefore also be carried out by third parties to provide safeguards against use for other purposes.

An active fairness approach does not mean, however, that France would commit itself to a policy of quotas for each of the 25 discrimination criteria. It is not up to the algorithms to define the right balance between each group. The risk of crystallizing society and reducing it to a competition between different groups would be too great. Rather, it is a matter of giving public and private actors the means to detect bias and discrimination.

- **Proposal #5: Make public test databases available to enable companies to assess biases within their methodology**

An active fairness approach requires a database comprising variables normally protected under the 25 discrimination criteria. In some cases, it will not be possible for companies to set up these test databases because either individuals will not agree to have their data collected despite the safeguards (impact analysis, collection on a small sample, for test purposes only) or it is not desired that companies collect this type of information. In these cases, the State, or an independent actor, could take charge of setting up such databases.

One case of use mentioned by many specialists concerns facial recognition. Developers do not have a database representative of the French population to test for the absence of bias in their algorithms. There is no way to verify that the algorithms will work correctly on the entire population residing in France. This is true for French developers, but also for those importing into France algorithms that have learned in the United States or China.

We recommend that public databases including information on some of the 25 protected criteria be made available for precise uses. This should be limited to specific cases such as facial recognition (gender, etc.) or credit risk assessment by banks (income history, gender, socio-professional category).

Such databases would be used to test the methodology, before or after learning on collected data. They would make it possible to verify that a protected variable had not been "rediscovered" by the algorithm as could be feared in the case of automobile insurance[50].

In the United States, the work of the National Institute of Standards and Technology (NIST) has established a benchmark. It provides databases to evaluate biometric technologies in terms of both performance and absence of bias. The U.S. government has granted this public institute access to large quantities of confidential data. Its scientific rigor is recognized in both the metrics and the protocols. In France, the Laboratoire National de Métrologie et d'Essais (LNE, National Laboratory of Metrology and Testing) could take on this mission. The LNE, under the supervision of the Ministry of Economy, has the mission to assist public authorities and economical actors in the elaboration of testing methods. It is responsible for a national challenge, designed to evaluate

50 An automobile insurance algorithm that does not have access to gender data could deliver higher premiums for red cars. This is indeed a good approximate measurement of gender, since vehicles of this color are more often owned by men (higher risk of accident).

agricultural robots, and has already made partnerships to develop evaluation methods towards automatic speech transcription algorithms.

The aim is not to publish databases containing highly confidential information on thousands of French citizens, especially since complete anonymization is increasingly seen as impossible. Access to these databases could be tightly controlled, as is the case for access to the statistical databases of the National Institute of Statistics and Economic Studies (INSEE). The possibility of using "synthetic" databases must also be explored. This would allow for using noisy data, i.e. data that are modified to publish databases with no real profile, while keeping the initial statistical properties intact.

Being the reference database implies great responsibilities, as the example of ImageNet has shown. It will be imperative to scrutinize these databases and to exact the highest standards for them.

D. Evaluate high-risk algorithms to limit their impact

We believe that it is necessary not to consider all algorithms in the same way. Just as there is no regulation for all machines (only the most dangerous ones are regulated), it is the high-risk algorithms that must attract the attention of public authorities, companies and civil society.

This is also the approach suggested by the European Commission in its draft White Paper on Artificial Intelligence published in January 2020[51], which suggests that algorithm developers and regulators should focus their efforts on high-risk algorithms.

But how is high risk defined? The European Commission mentions two cumulative criteria: an algorithm must (I) concern a sector defined as high-risk (health care, transport, policing, other) and (II) "produce legal impact for the individual or the legal entity or pose risk of injury, death or significant material damage." Beyond this definition, which is limited to legal and material impacts, three factors are important in our opinion: denial of access to essential services[52], infringement of the safety of individuals[53], and the restriction of fundamental rights[54]. Algorithms resulting in decisions that generate one or more of these types of impact can be qualified as critical.

51 European Commission's draft White Paper on Artificial Intelligence.
52 Bank domiciliation, social security, employment, etc.
53 Physical as well as moral injury.
54 Freedom of movement, right to demonstrate, etc.

Examples of high-risk algorithms

In the automotive industry, certain very basic algorithms are used to adjust the temperature of the vehicle or the inclination of the driver's seat. Conversely, other algorithms define the behavior of an autonomous vehicle. A bias in the second instance would have a very strong impact, contrary to a bias in the second. Both industrialists and regulators in the sector have taken the full measure of this issue. Standard ISO26262, which concerns critical systems, defined by the International Organisation for Standardisation, specifically regulates algorithms that have an impact on passenger safety, not the others. This standard is, moreover, currently undergoing modernization work to adapt it to artificial intelligence technologies. High-risk algorithms in this area are primarily those that affect the safety of passengers and passers-by.

In the banking sector, some algorithms will carry out targeted advertising, while others will evaluate for each client his/her risk and therefore eligibility for a bank loan. These algorithms are naturally optimized to reduce the overall portfolio risk. An improvement to the code will be applied as soon as it improves the overall risk profile, even if this implies lower accuracy levels for some individuals. The consequences on the lives of these individuals are potentially significant as their access to essential services is reduced.

In the health sector, some algorithms will be used to detect a small fracture on an X-ray, while others will determine from biological data whether the patient shows a recurrence of cancer. One of these two will involve the patient's vital prognosis, and significant bias would be much more critical in this case. Here again, the US Federal Drug Administration (FDA) has taken the full measure of this issue. In its discussion paper on the regulatory framework to be applied to medical devices comprising self-learning software (SaMD), it differentiates algorithms according to their risk (criticality for the patient and impact of the algorithm on the clinical decision). If self-learning changes the risk for the patient, then the device should be able to prove that it is still accurate. Otherwise, it will be able to evolve as it wishes. The analysis grid proposed by the Haute Autorité de Santé (High Authority on Health) to evaluate medical devices incorporating artificial intelligence also includes a question on the impact that biases in data collection could have.

.../...

In the area of policing, the deployment of facial recognition algorithms could, in the presence of bias, restrict the fundamental rights of certain groups of citizens. Both their right to demonstrate and the presumption of innocence would be affected and limited with respect to other groups. Conversely, algorithms for the automated translation of foreign languages involve, in theory, less risk, as far as established biases are concerned

We recommend that each stakeholder developing algorithms set up specific requirements for these high-risk algorithms. The State will also have a role in stimulating the development of labels and certifications for these algorithms in order to promote and strengthen confidence in their operation, as well as in developing its own capability to audit these algorithms as a last resort.

- **Proposal #6: Implement more stringent requirements for high-risk algorithms**

The need to focus responses to algorithmic biases on systems with the greatest impact leads us to define a specific associated framework. When an algorithm presents a risk of significant impact, two measures become essential: transparency, and the right of appeal. We recommend that public and private actors implementing high-risk algorithms make these two measures effective.

Transparency consists in disclosing, whether to the users of the algorithms or to a trusted third party, information about its purpose, methodology, data collected and number of citizens impacted. The need for transparency is proportional to the impact: the demands on food have thus grown as the content and origin of our foodstuffs has become more and more vital for good health.

Food available for purchase in a supermarket is accompanied by a lot of information on the product's expiry date, calorific content, and ingredients. This information is necessary to reassure consumers about the quality of what they are eating. Even if the consumers do not read the information, its mere presence reassures because it guarantees that the manufacturer has nothing to hide.

The success of initiatives such as Nutriscore or Yuka, which make it possible to assess the nutritional quality of food purchased in a supermarket with just one click, demonstrates the appetite of the French for clear, simple and quantified information on the quality of the products they buy.

In the field of health, the bioethics law currently under review requires that algorithms used for medical diagnosis or treatment include an explanatory leaflet. Here this law merely reiterates a common sense principle: the greater the impact of the object, the more guarantees are necessary concerning its operation.

We believe it is essential to multiply this type of initiative for high-risk algorithms by focusing transparency efforts on the nature and quality of the data used and on the objectives of the algorithms. Biases come above all from (I) the data and (II) the choice of the objectives that the algorithm must optimize (see Part II). It is therefore on these elements that we must focus.

If we take the example of the Yuka app, our proposal amounts to suggesting the publication of two types of information. First, a list of the type of data used to train the algorithm in assessing the nutritional quality of foods. This list could be supplemented by an indicator on the quality of these data (homogeneous and representative data, fairly large sample, etc.), as well as information on the conditions of data acquisition and management. Indeed, it is necessary to address the quality of the underlying data (the ingredients) rather than the code of the algorithm (the recipe), more visible but often less biased.

And secondly, a publication of the objectives that the algorithm must achieve. Is it trained to minimize the calorific value of recommended foods, to place emphasis on certain types of diets (for example, salt-free), or to promote the value of certain types of foods (organic, for example)? The nature of the objectives is an essential element to understand where the algorithm wants to lead us and to identify any possible biases.

In this example, the transparency is for the benefit of the consumer. Nevertheless, transparency with respect to a trusted third party, public body, research laboratory or other, is quite conceivable and would allow for protecting the professional secrecy surrounding this strategic information.

The second measure within this specific framework concerns the right of appeal. Where an important decision is taken or influenced by an algorithm, there must be a corresponding right to challenge this decision. This principle, enshrined in French and European law for automated, public or high-impact decisions, is an absolutely essential element for the development of fair algorithms. A legitimate criticism of Parcoursup (the national platform for pre-registration in the first year of higher education in France) is, notably, the fact that it has not clearly opened up a means of appeal inside the process. This second measure is likely to strengthen confidence in such algorithms.

The right of appeal against automated decision-making system could be largely based on the right of appeal as it exists today, and in particular on the institution of the French Ombudsman. An employee who suspects, for example, that a discriminatory decision has been taken against him or her, can refer the matter to the Ombudsman. It is then no longer up to the employee to prove that discrimination has occurred, but rather to the company under attack to prove that there has been no such discrimination. A right of appeal where the burden of proof lies with the developer of the algorithm must necessarily go hand in hand with the authorization to collect certain protected variables for testing purposes (see proposal #4 on active fairness).

This right is not always possible. In the case of automated driving algorithms, an appeal procedure will be of little use in the event of an accident. On the other hand, algorithm redundancy, i.e. the presence of a second algorithm that would analyze the first decision and possibly contradict it, is a solution that should be promoted, especially when personal safety is at stake. Thus, two automated driving algorithms could, when giving different orders, alert the driver to a limit in the algorithm.

This specific framework does not require a major law on algorithmic bias common to all sectors of activity. We believe that the adoption of good practices by the companies and administrations concerned, the use of existing mechanisms such as the Ombudsman, and the addition of provisions in sectoral legislation on a case-by-case basis are sufficient.

- **Proposal #7: Support the emergence of labels to strengthen the confidence of citizens in critical uses, and accelerate the dissemination of beneficial algorithms**

A commitment to transparency with respect to consumers or corporate clients can be difficult to implement. Indeed, information such as the objectives of the algorithms, the variables used as inputs, or the weights between the different variables, can take on a strong strategic character. This is particularly true in areas such as banking, insurance or automated driving, where the algorithmic of the algorithms will differentiate competitors in the long term.

Moreover, transparency is very time-consuming for the person who has to analyze the elements provided and form an opinion. The consumer reading household appliance manuals and the SME analyzing the specifications of a newly acquired industrial machine have to invest a considerable amount of time to be reassured about the quality and correct operation of their purchase.

The industrial sector has since long developed quality and safety labels that guarantee compliance with a certain level of requirements for both individual and corporate buyers. For the most critical uses, these are mandatory certifications, based on the object's compliance with safety cases[55]. This is notably the standard in the automotive, aviation and health sectors. For less critical uses, these may simply be (optional) labels, which give the buyer an indication on the product's characteristics (e.g. the AB organic farming label in France).

We recommend transposing this logic of industrial quality to algorithms by promoting the emergence of specific labels and certifications. The process of defining a certification being slow and restrictive in relation to the rate of evolution of the technology, it is the emergence of labels that should be prioritized in the short term.

Given the difficulty of defining standards for biases, it seems illusory to define a label guaranteeing a bias-free algorithm. Nevertheless, labels could focus on the auditability of algorithms, on data quality, or on the presence of a bias risk assessment process within the company.

The Fair Data Use label is an example of the type of initiative that should be encouraged in each sector, particularly for high-risk algorithms. This label[56] is obtained after an audit of the algorithms to guarantee the absence of discrimination, and compliance with the GDPR and the company's CSR rules. This label, delivered for one year, consists of an evaluation of the targeted algorithm by an auditing algorithm, which analyzes precise criteria, such as the presence of sensitive variables, transparency or loyalty.

The development of labels would also make it possible to develop skills and methods for auditing algorithms within the French ecosystem. As an example, the data augmentation method allows to test the independence of an algorithm vis-à-vis specific variables. It consists in generating artificial data, for example by creating individuals that are identical to the ones in the database in all aspects but gender. We could therefore check that, with equal profiles, a woman and a man are treated in the same way by the algorithm. Other methods allow for inferring variables that are explicitly excluded from the database, but which can nevertheless be captured by the algorithm. Specialists in these methods could work to build the foundations of an algorithmic testing industry.

55 Definition of a situation in which the object is required to meet a number of safety criteria. For example, crash tests for cars or clinical studies for drugs.
56 Website of the label *"Fair data use"* by Maathics.

One difficulty that labels and certifications will encounter when extended to the domain of algorithmic bias is the frequency of evolution of source codes and learning data. Labels focusing on the team developing an algorithm (its internal processes, practices, composition) rather than on the algorithms themselves may help to overcome this difficulty.

• Proposal #8: Develop the capability to audit high-risk algorithms

In the justice sector, algorithms that seek to predict recidivism, and therefore determine bail, are not inherently less efficient than the judges. They also make choices that can be biased, and are likewise subject to the same contradictory injunctions: on equality between groups (they must be wrong as many times for young people as for elderly people) and on equality between individuals (two identical profiles must be treated independently of skin color.) But the algorithm generates a new problem when it is developed by a private company, when it is protected by trade secrets, when its learning data is not available, and when it cannot be audited. Where no appeal is possible against the decision that has been taken, the justice system is undermined.

We are convinced that our training recommendations and best practices will promote the development of less biased algorithms. The emergence of labels will make it possible to disseminate these good practices, and to promote virtuous developers of high-risk algorithms. In cases where bias nevertheless occurs, and with a high risk, appeal options will give individuals or civil society the possibility to avoid being subject to bias without being able to take action.

The fact remains that in some extremely problematic or sensitive cases, it must be possible to audit algorithms, their data and their operation. These audits of high-risk algorithms could be carried out by third parties, along the lines of auditors of corporate financial accounts, or by the State.

The cases of Parcoursup, facial recognition by the police, or predictive algorithms in the American justice system, constitute multiple proof that the State concentrates many of the high-impact uses. However, its expertise and auditing capability, as a regulator and buyer, is today both limited and fragmented. Seven State authorities and departments are potentially competent for the audiovisual and communication sectors[57]. The creation of a pole of expertise capable of auditing algorithms, initiated in 2019, is thus good news.

57 French Competition Authority, CSA, CNIL, ARCEP, DGE, DGCCRF, Directorate General of Media and Cultural Industries.

The French State sets up a center for algorithmic expertise

The idea of merging several of these regulators to create a single digital regulator is often put forward: CSA and CNIL, CNIL and ARCEP, CSA and ARCEP. Without wishing to comment on this subject, we note that while audiovisual and communication are at the forefront of the sectors transformed by algorithms, they are not the only ones.

The cross-cutting nature of the digital players and the omnipresence of algorithms in the future require the creation of digital expertise capabilities within the State, mobilized around different missions. The diversity and importance of the projects would facilitate the recruitment of technical profiles that are currently in great demand.

The bill on audiovisual communication and cultural sovereignty in the digital age, tabled in December 2019, provides for the creation of a "digital expertise center" of 20 people at the Directorate General for Enterprise (DGE) of the French Ministry of the Economy. This project responds to a need that goes far beyond the audiovisual field alone.

Algorithms are constantly evolving, either because they are improved or because they learn based on feedback loop. Unfortunately, a one-off audit of an algorithm can quickly become obsolete. The study by AlgoTransparency on the bias of YouTube's algorithm in favor of radical content can be disproved six months later because YouTube takes criticism into account on a regular basis.

In cases where an algorithm can have a very high impact, it is desirable to have continuous assurance on the absence of bias. A trusted third party such as auditing companies or the State could choose to implement specifically digital control[58]. Rather than having a paper audit once a year, the third party could retrieve either test results or data to audit the algorithm via secure APIs. It could thus continuously check that the algorithms comply with a number of criteria, notably in terms of bias.

58 Grossman N., *Regulation, the internet way*, Data-smart city solutions, ASH Center, Harvard Kennedy School, April 8, 2015.

CONCLUSION

Algorithmic biases have emerged in the American debate on discrimination in the digital age. Particularly striking cases in the field of justice, recruitment or access to financial services have marked public opinion and raised awareness on this subject among the general public, researchers, companies and public authorities.

However, it is not enough to think that this awareness on the other side of the Atlantic protects us from the risks associated with these biases. The definition of fairness, of the behavior that the algorithm should adopt to ensure fair decisions, is not a universal concept. France can and must have its own doctrine with regard to algorithmic bias, based on its culture and its political and legal history

This French approach will have to reconcile, on the one hand, the fact that Europe is lagging behind in terms of digital technology, something that should not be further exacerbated by regulation against algorithmic bias, and, on the other hand, the very real risk of social destabilization.

It will also have to take into account the necessary flexibility needed by all stakeholders concerned to adapt the objectives of the algorithms and the associated requirements of fairness and performance to each context.

Finally, this approach will have to integrate the exceptional potential for reducing discrimination that algorithms represent. The issue at stake is, obviously, determining whether algorithms are biased, but especially whether they are more biased than the human beings they replace or assist. Without underestimating the risks inherent to algorithms such as lack of transparency or the ability to multiply and standardize a biased decision, the possibility to reduce discrimination thanks to algorithms is real.

We are convinced that it is too early to propose a law on algorithmic bias or the ex ante control of algorithms by the State. The legal framework on discrimination and digital technologies already offers many solutions and its implementation should be strengthened. Moreover, the State, having great difficulty today to fully enforce the European General Data Protection Regulation, would also find it hard to verify such a large number of algorithms, which are becoming increasingly complex.

Given the still moderate extent of the phenomenon in Europe, we believe that this French doctrine on algorithmic bias should be based on three pillars. Firstly, an essential training effort to ensure that all the actors and stakeholders in the algorithm value chain are aware of the risks associated with the deployment of algorithms.

Secondly, the creation of capabilities for public and private actors implementing algorithms to test them for possible biases.

And thirdly, dissociation of the most sensitive algorithms from all others in order to make them subject to a specific framework. These are algorithms that infringe fundamental rights, jeopardize the physical or psychological safety of individuals, and restrict access to essential services. For these algorithms, rights of appeal, transparency, labeling and third party auditing are steps that will necessarily have to be included in their development.

The legislative and regulatory initiative on algorithms and artificial intelligence envisaged by the new European Commission will therefore represent an essential step in the development of a French and European doctrine with respect to the risks of algorithmic bias.

ACKNOWLEDGMENTS

Institut Montaigne would like to thank the following persons for their contribution to this work.

Task Force - Chairpersons

- **Anne Bouverot,** Chairman of the Board at Technicolor and Chairman of the Fondation Abeona
- **Thierry Delaporte,** Chief Operating Officer, Capgemini

Task Force - Rapporteurs

- **Arno Amabile,** Engineer, Corps des Mines
- **Théophile Lenoir,** Head of Digital Program, Institut Montaigne
- **Tanya Perelmuter,** Director of Strategy & Partnerships, Fondation Abeona (General Rapporteur)
- **Basile Thodoroff,** Engineer, Corps des Mines

Task Force - Members

- **Gilles Babinet,** Digital Advisor, Institut Montaigne
- **Ingrid Bianchi,** Founder/Director, Diversity Source Manager
- **David Bounie,** Director of the Social and Economic Sciences Department, Télécom Paris
- **Dominique Cardon,** Director, Médialab, Sciences Po
- **Anna Choury,** Advanced Data Analytics Manager, Airbus
- **Stephan Clémençon,** Professor and Researcher, Télécom Paris
- **Marcin Detyniecki,** Head of Research and Development & Group Chief Data Scientist, AXA
- **Dominique Latourelle,** Head of RTB, iProspect
- **Sébastien Massart,** Director of Strategy, Dassault Systèmes
- **Bernard Ourghanlian,** Chief Technology Officer and Chief Security Officer, Microsoft France
- **Guillemette Picard,** Chief Health Officer, Nabla
- **Christian de Sainte Marie,** Director, Center of Advanced Studies, IBM France
- **François Sillion,** Director, Advanced Technologies Center Paris, Uber
- **Serge Uzan,** Vice-President, Conseil national de l'ordre des médecins

As well as:

- **Joan Elbaz,** Policy Officer Assistant, Institut Montaigne
- **Margaux Tellier,** Policy Officer Assistant, Institut Montaigne
- **Julie Van Muylders,** Policy Officer Assistant, Institut Montaigne

Interviewed persons in the making of this work:

- **Éric Adrian,** General Manager, UiPath France
- **Prabhat Agarwal,** Deputy Head of Unit E-Commerce and Platforms, DG Connect, European Commission
- **Sacha Alanoca,** Senior AI Policy Researcher & Head of Community Development, The Future Society
- **Christine Bargain,** Director of Social Responsibility from 2011 to 2018, Groupe La Poste
- **Marie Beaurepaire,** Project Officer, Afmd
- **Bertrand Braunschweig,** Coordination Director of the National Research Program on Artificial Intelligence
- **Alexandre Briot,** Artificial Intelligence Team Leader, Valeo
- **Clément Calauzènes,** Senior Staff Research Lead, Criteo AI Lab
- **Laurent Cervoni,** Director of Artificial Intelligence, Talan
- **Florence Chafiol,** Partner, August Debouzy
- **Guillaume Chaslot,** Mozilla Fellow and Founder, Algotransparency
- **Raja Chatila,** Intelligence, Robotics and Ethics Professor, and Member of the High-Level Expert Group on Artificial Intelligence, European Commission
- **Bertrand Cocagne,** Director of Innovation and Technologies Lending & Leasing, Linedata Services
- **Guillaume De Saint Marc,** Senior Director, Chief Technology and Architecture Office, Cisco
- **Marie-Laure Denis,** Chairperson, CNIL
- **Christel Fiorina,** Coordinator of the Economic Part of the National Strategy on Artificial Intelligence
- **Marie-Anne Frison-Roche,** Professor, Sciences Po
- **Vincent Grari,** Research Data Scientist, AXA
- **Arthur Guillon,** Senior Machine Learning Engineer, easyRECrue
- **Nicolas Kanhonou,** Director, Promotion of Equality and Access to Rights, Défenseur des droits
- **Djamil Kemal,** co-CEO, Goshaba
- **Yann Le Biannic,** Data Science Chief Expert, SAP
- **Agnès Malgouyres,** Head of Artificial Intelligence, Siemens Healthineers France
- **Stéphane Mallat,** Professor, Collège de France
- **Sébastien Mamessier,** Senior Research Engineer, Uber
- **Claire Mathieu,** Director of Research, CNRS
- **Marc Mézard,** Director, ENS
- **Nicolas Miailhe,** Co-founder and Chairperson, The Future Society
- **Christophe Montagnon,** Director of Organisation, Computer Systems and Quality, Randstad France

- **Christelle Moreux,** Chief Legal Officer, Siemens Healthcare
- **François Nédey,** Head of Technical Unit and Products, Member of the Board, Allianz
- **Bertrand Pailhès,** Coordinator of the French Strategy in Artificial Intelligence until November 2019, Director of Technologies and Innovation, CNIL
- **Cédric Puel,** Head of Data and Analytics, BNP Paribas Retail Banking and Services
- **Pete Rai,** Principal Engineer in the Chief Technology and Architecture Office, Cisco
- **Boris Ruf,** Research Data Scientist, AXA
- **Bruno Sportisse,** Chairperson and General Director, Inria
- **Pierre Vaysse,** Head of Retail P&C and Pricing, Allianz France
- **Renaud Vedel,** Ministry Coordinator in the Field of Artificial Intelligence, Ministry of the Interior
- **Fernanda Viégas,** Co-lead, PAIR Initiative, Google

The views set out in this report do not reflect the opinions of the people previously mentioned nor the institutions they represent.

OUR PREVIOUS PUBLICATIONS

- Retraites : pour un régime équilibré (mars 2020)
- Space: Will Europe Awaken? (Février 2020)
- Données personnelles : comment gagner la bataille ? (décembre 2019)
- Transition énergétique : faisons jouer nos réseaux (décembre 2019)
- Religion au travail : croire au dialogue - Baromètre du Fait Religieux Entreprise 2019 (novembre 2019)
- Taxes de production : préservons les entreprises dans les territoires (octobre 2019)
- Médicaments innovants : prévenir pour mieux guérir (septembre 2019)
- Rénovation énergétique : chantier accessible à tous (juillet 2019)
- Agir pour la parité : performance à la clé (juillet 2019)
- Pour réussir la transition énergétique (juin 2019)
- Europe-Afrique : partenaires particuliers (juin 2019)
- Media polarization « à la française » ? Comparing the French and American ecosystems (mai 2019)
- L'Europe et la 5G : le cas Huawei (partie 2, mai 2019)
- L'Europe et la 5G : passons la cinquième ! (partie 1, mai 2019)
- Système de santé : soyez consultés ! (avril 2019)
- Travailleurs des plateformes : liberté oui, protection aussi (avril 2019)
- Action publique : pourquoi faire compliqué quand on peut faire simple (mars 2019)
- La France en morceaux : baromètre des Territoires 2019 (février 2019)
- Énergie solaire en Afrique : un avenir rayonnant ? (février 2019)
- IA et emploi en santé : quoi de neuf docteur ? (janvier 2019)
- Cybermenace : avis de tempête (novembre 2018)
- Partenariat franco-britannique de défense et de sécurité : améliorer notre coopération (novembre 2018)
- Sauver le droit d'asile (octobre 2018)
- Industrie du futur, prêts, partez ! (septembre 2018)
- La fabrique de l'islamisme (septembre 2018)
- Protection sociale : une mise à jour vitale (mars 2018)
- Innovation en santé : soignons nos talents (mars 2018)
- Travail en prison : préparer (vraiment) l'après (février 2018)
- ETI : taille intermédiaire, gros potentiel (janvier 2018)
- Réforme de la formation professionnelle : allons jusqu'au bout ! (janvier 2018)
- Espace : l'Europe contre-attaque ? (décembre 2017)
- Justice : faites entrer le numérique (novembre 2017)
- Apprentissage : les trois clés d'une véritable transformation (octobre 2017)
- Prêts pour l'Afrique d'aujourd'hui ? (septembre 2017)

- Nouveau monde arabe, nouvelle « politique arabe » pour la France (août 2017)
- Enseignement supérieur et numérique : connectez-vous ! (juin 2017)
- Syrie : en finir avec une guerre sans fin (juin 2017)
- Énergie : priorité au climat! (juin 2017)
- Quelle place pour la voiture demain? (mai 2017)
- Sécurité nationale : quels moyens pour quelles priorités? (avril 2017)
- Tourisme en France : cliquez ici pour rafraîchir (mars 2017)
- L'Europe dont nous avons besoin (mars 2017)
- Dernière chance pour le paritarisme de gestion (mars 2017)
- L'impossible État actionnaire? (janvier 2017)
- Un capital emploi formation pour tous (janvier 2017)
- Économie circulaire, réconcilier croissance et environnement (novembre 2016)
- Traité transatlantique : pourquoi persévérer (octobre 2016)
- Un islam français est possible (septembre 2016)
- Refonder la sécurité nationale (septembre 2016)
- Bremain ou Brexit : Europe, prépare ton avenir! (juin 2016)
- Réanimer le système de santé - Propositions pour 2017 (juin 2016)
- Nucléaire : l'heure des choix (juin 2016)
- Un autre droit du travail est possible (mai 2016)
- Les primaires pour les Nuls (avril 2016)
- Le numérique pour réussir dès l'école primaire (mars 2016)
- Retraites : pour une réforme durable (février 2016)
- Décentralisation : sortons de la confusion / Repenser l'action publique dans les territoires (janvier 2016)
- Terreur dans l'Hexagone (décembre 2015)
- Climat et entreprises : de la mobilisation à l'action / Sept propositions pour préparer l'après-COP21 (novembre 2015)
- Discriminations religieuses à l'embauche : une réalité (octobre 2015)
- Pour en finir avec le chômage (septembre 2015)
- Sauver le dialogue social (septembre 2015)
- Politique du logement : faire sauter les verrous (juillet 2015)
- Faire du bien vieillir un projet de société (juin 2015)
- Dépense publique : le temps de l'action (mai 2015)
- Apprentissage : un vaccin contre le chômage des jeunes (mai 2015)
- Big Data et objets connectés. Faire de la France un champion de la révolution numérique (avril 2015)
- Université : pour une nouvelle ambition (avril 2015)
- Rallumer la télévision : 10 propositions pour faire rayonner l'audiovisuel français (février 2015)
- Marché du travail : la grande fracture (février 2015)

- Concilier efficacité économique et démocratie : l'exemple mutualiste (décembre 2014)
- Résidences Seniors : une alternative à développer (décembre 2014)
- Business schools : rester des champions dans la compétition internationale (novembre 2014)
- Prévention des maladies psychiatriques : pour en finir avec le retard français (octobre 2014)
- Temps de travail : mettre fin aux blocages (octobre 2014)
- Réforme de la formation professionnelle : entre avancées, occasions manquées et pari financier (septembre 2014)
- Dix ans de politiques de diversité : quel bilan ? (septembre 2014)
- Et la confiance, bordel ? (août 2014)
- Gaz de schiste : comment avancer (juillet 2014)
- Pour une véritable politique publique du renseignement (juillet 2014)
- Rester le leader mondial du tourisme, un enjeu vital pour la France (juin 2014)
- 1 151 milliards d'euros de dépenses publiques : quels résultats ? (février 2014)
- Comment renforcer l'Europe politique (janvier 2014)
- Améliorer l'équité et l'efficacité de l'assurance-chômage (décembre 2013)
- Santé : faire le pari de l'innovation (décembre 2013)
- Afrique-France : mettre en œuvre le co-développement
  Contribution au XXVIe sommet Afrique-France (décembre 2013)
- Chômage : inverser la courbe (octobre 2013)
- Mettre la fiscalité au service de la croissance (septembre 2013)
- Vive le long terme ! Les entreprises familiales au service de la croissance et de l'emploi (septembre 2013)
- Habitat : pour une transition énergétique ambitieuse (septembre 2013)
- Commerce extérieur : refuser le déclin
  Propositions pour renforcer notre présence dans les échanges internationaux (juillet 2013)
- Pour des logements sobres en consommation d'énergie (juillet 2013)
- 10 propositions pour refonder le patronat (juin 2013)
- Accès aux soins : en finir avec la fracture territoriale (mai 2013)
- Nouvelle réglementation européenne des agences de notation : quels bénéfices attendre ? (avril 2013)
- Remettre la formation professionnelle au service de l'emploi et de la compétitivité (mars 2013)
- Faire vivre la promesse laïque (mars 2013)
- Pour un « New Deal » numérique (février 2013)
- Intérêt général : que peut l'entreprise ? (janvier 2013)

- Redonner sens et efficacité à la dépense publique 15 propositions pour 60 milliards d'économies (décembre 2012)
- Les juges et l'économie : une défiance française ? (décembre 2012)
- Restaurer la compétitivité de l'économie française (novembre 2012)
- Faire de la transition énergétique un levier de compétitivité (novembre 2012)
- Réformer la mise en examen Un impératif pour renforcer l'État de droit (novembre 2012)
- Transport de voyageurs : comment réformer un modèle à bout de souffle ? (novembre 2012)
- Comment concilier régulation financière et croissance : 20 propositions (novembre 2012)
- Taxe professionnelle et finances locales : premier pas vers une réforme globale ? (septembre 2012)
- Remettre la notation financière à sa juste place (juillet 2012)
- Réformer par temps de crise (mai 2012)
- Insatisfaction au travail : sortir de l'exception française (avril 2012)
- Vademecum 2007 – 2012 : Objectif Croissance (mars 2012)
- Financement des entreprises : propositions pour la présidentielle (mars 2012)
- Une fiscalité au service de la « social compétitivité » (mars 2012)
- La France au miroir de l'Italie (février 2012)
- Pour des réseaux électriques intelligents (février 2012)
- Un CDI pour tous (novembre 2011)
- Repenser la politique familiale (octobre 2011)
- Formation professionnelle : pour en finir avec les réformes inabouties (octobre 2011)
- Banlieue de la République (septembre 2011)
- De la naissance à la croissance : comment développer nos PME (juin 2011)
- Reconstruire le dialogue social (juin 2011)
- Adapter la formation des ingénieurs à la mondialisation (février 2011)
- « Vous avez le droit de garder le silence… » Comment réformer la garde à vue (décembre 2010)
- Gone for Good ? Partis pour de bon ? Les expatriés de l'enseignement supérieur français aux États-Unis (novembre 2010)
- 15 propositions pour l'emploi des jeunes et des seniors (septembre 2010)
- Afrique - France. Réinventer le co-développement (juin 2010)
- Vaincre l'échec à l'école primaire (avril 2010)
- Pour un Eurobond. Une stratégie coordonnée pour sortir de la crise (février 2010)
- Réforme des retraites : vers un big-bang ? (mai 2009)
- Mesurer la qualité des soins (février 2009)

- Ouvrir la politique à la diversité (janvier 2009)
- Engager le citoyen dans la vie associative (novembre 2008)
- Comment rendre la prison (enfin) utile (septembre 2008)
- Infrastructures de transport : lesquelles bâtir, comment les choisir ? (juillet 2008)
- HLM, parc privé
  Deux pistes pour que tous aient un toit (juin 2008)
- Comment communiquer la réforme (mai 2008)
- Après le Japon, la France…
  Faire du vieillissement un moteur de croissance (décembre 2007)
- Au nom de l'Islam… Quel dialogue avec les minorités musulmanes en Europe ?
  (septembre 2007)
- L'exemple inattendu des Vets
  Comment ressusciter un système public de santé (juin 2007)
- Vademecum 2007-2012
  Moderniser la France (mai 2007)
- Après Erasmus, Amicus
  Pour un service civique universel européen (avril 2007)
- Quelle politique de l'énergie pour l'Union européenne ? (mars 2007)
- Sortir de l'immobilité sociale à la française (novembre 2006)
- Avoir des leaders dans la compétition universitaire mondiale (octobre 2006)
- Comment sauver la presse quotidienne d'information (août 2006)
- Pourquoi nos PME ne grandissent pas (juillet 2006)
- Mondialisation : réconcilier la France avec la compétitivité (juin 2006)
- TVA, CSG, IR, cotisations…
  Comment financer la protection sociale (mai 2006)
- Pauvreté, exclusion : ce que peut faire l'entreprise (février 2006)
- Ouvrir les grandes écoles à la diversité (janvier 2006)
- Immobilier de l'État : quoi vendre, pourquoi, comment (décembre 2005)
- 15 pistes (parmi d'autres…) pour moderniser la sphère publique (novembre 2005)
- Ambition pour l'agriculture, libertés pour les agriculteurs (juillet 2005)
- Hôpital : le modèle invisible (juin 2005)
- Un Contrôleur général pour les Finances publiques (février 2005)
- Les oubliés de l'égalité des chances (janvier 2004 - Réédition septembre 2005)

For previous publications, see our website:
www.institutmontaigne.org/en

The information and views set out in this report are those of Institut Montaigne and do not necessarily reflect the opinions of the people and institutions mentioned above.

INSTITUT
MONTAIGNE

ABB FRANCE
ABBVIE
ACCURACY
ACTIVEO
ADIT
ADVANCY
AIR FRANCE - KLM
AIR LIQUIDE
AIRBUS
ALLEN & OVERY
ALLIANZ
ALVAREZ & MARSAL FRANCE
AMAZON WEB SERVICES
AMBER CAPITAL
AMUNDI
ARCHERY STRATEGY CONSULTING
ARCHIMED
ARDIAN
ASTORG
ASTRAZENECA
AUGUST DEBOUZY
AVRIL
AXA
BAKER & MCKENZIE
BANK OF AMERICA MERRILL LYNCH
BEARINGPOINT
BESSÉ
BNP PARIBAS
BOLLORÉ
BOUGARTCHEV MOYNE ASSOCIÉS
BOUYGUES
BROUSSE VERGEZ
BRUNSWICK
CAISSE DES DÉPÔTS
CANDRIAM
CAPGEMINI
CAPITAL GROUP
CAREIT
CARREFOUR
CASINO
CHAÎNE THERMALE DU SOLEIL
CHUBB
CIS
CISCO SYSTEMS FRANCE
CMA CGM
CNP ASSURANCES
COHEN AMIR-ASLANI
COMPAGNIE PLASTIC OMNIUM
CONSEIL SUPÉRIEUR DU NOTARIAT
CORREZE & ZAMBEZE
CRÉDIT AGRICOLE
CRÉDIT FONCIER DE FRANCE
D'ANGELIN &CO.LTD
DASSAULT SYSTÈMES
DE PARDIEU BROCAS MAFFEI
DENTSU AEGIS NETWORK
DRIVE INNOVATION INSIGHT - DII
EDF
EDHEC BUSINESS SCHOOL
EDWARDS LIFESCIENCES
ELSAN
ENEDIS
ENGIE
EQUANCY
ESL & NETWORK
ETHIQUE & DÉVELOPPEMENT
EURAZEO
EUROGROUP CONSULTING
EUROSTAR
FIVES
FONCIA GROUPE
FONCIÈRE INEA
GALILEO GLOBAL EDUCATION
GETLINK
GIC PRIVATE LIMITED
GIDE LOYRETTE NOUEL
GOOGLE
GRAS SAVOYE
GROUPAMA
GROUPE EDMOND DE ROTHSCHILD
GROUPE M6
HAMEUR ET CIE
HENNER
HSBC FRANCE
IBM FRANCE
IFPASS
ING BANK FRANCE
INSEEC
INTERNATIONAL SOS
INTERPARFUMS
IONIS EDUCATION GROUP
ISRP
JEANTET ASSOCIÉS
KANTAR
KATALYSE
KEARNEY
KPMG S.A.
LA BANQUE POSTALE

INSTITUT
MONTAIGNE

LA PARISIENNE ASSURANCES
LAZARD FRÈRES
LINEDATA SERVICES
LIR
LIVANOVA
L'ORÉAL
LOXAM
LVMH
M.CHARRAIRE
MACSF
MALAKOFF MÉDÉRIC
MAREMMA
MAZARS
MCKINSEY & COMPANY FRANCE
MÉDIA-PARTICIPATIONS
MEDIOBANCA
MERCER
MERIDIAM
MICHELIN
MICROSOFT FRANCE
MITSUBISHI FRANCE S.A.S
NATIXIS
NEHS
NESTLÉ
NEXITY
OBEA
ODDO BHF
ONEPOINT
ONDRA PARTNERS
ONET
OPTIGESTION
ORANGE
ORANO
ORTEC GROUPE
PAI PARTNERS
PRICEWATERHOUSECOOPERS
PRUDENTIA CAPITAL
RADIALL
RAISE
RAMSAY GÉNÉRALE DE SANTÉ
RANDSTAD
RATP
RELX GROUP
RENAULT
REXEL
RICOL LASTEYRIE CORPORATE FINANCE
RIVOLIER
ROCHE
ROLAND BERGER
ROTHSCHILD MARTIN MAUREL
SAFRAN
SANOFI
SAP FRANCE
SCHNEIDER ELECTRIC
SERVIER
SGS
SIA PARTNERS
SIACI SAINT HONORÉ
SIEMENS FRANCE
SIER CONSTRUCTEUR
SNCF
SNCF RÉSEAU
SODEXO
SOFINORD - ARMONIA
SOLVAY
SPRINKLR
SPVIE
STAN
SUEZ
TALAN
TECNET PARTICIPATIONS SARL
TEREGA
TETHYS
THE BOSTON CONSULTING GROUP
TILDER
TOTAL
TRANSDEV
UBER
UBS FRANCE
UIPATH
VEOLIA
VINCI
VIVENDI
VOYAGEURS DU MONDE
WAVESTONE
WAZE
WENDEL
WORDAPPEAL
WILLIS TOWERS WATSON

SUPPORT INSTITUT MONTAIGNE

INSTITUT
MONTAIGNE

Imprimé en France
Dépôt légal : mars 2020
ISSN : 1771-6756
Achevé d'imprimer en mars 2020

INSTITUT
MONTAIGNE

THERE IS NO DESIRE MORE NATURAL THAN THE DESIRE FOR KNOWLEDGE

Algorithms: Please Mind the Bias!

Algorithms help us all day long, by calculating the shortest route on our phones or automatically creating playlists with our favorite songs. If they are incredibly efficient, they can also be perceived as black boxes - making decisions out of our control. What happens if a recruiting algorithm systematically lets women or ethnic minorities aside? How to make sure these mistakes are acknowledged and corrected?

This report provides a French perspective on this issue, today essentially viewed through an American lens. It continues the study by Télécom Paris and Fondation Abeona, Algorithms: Bias, Discrimination and Fairness, published in 2019. Based on this technical background, and through the forty interviews conducted, our aim is to provide concrete solutions to limit potential abuses and increase trust in algorithms.

Institut Montaigne
59, rue La Boétie - 75008 Paris
Tél. +33 (0)1 53 89 05 60
www.institutmontaigne.org

ISSN 1771-6756
MARCH 2020