Can Everyday AI be Ethical?

Machine Learning Algorithm Fairness

Philippe Besse[1], Céline Castets-Renard[2], Aurélien Garivier[3] & Jean-Michel Loubes[4]

Abstract

Combining big data and machine learning algorithms, the power of automatic decision tools induces as much hope as fear. Many recently enacted European legislation (GDPR) and French laws attempt to regulate the use of these tools. Leaving aside the well-identified problems of data confidentiality and impediments to competition, we focus on the risks of discrimination, the problems of transparency and the quality of algorithmic decisions. The detailed perspective of the legal texts, faced with the complexity and opacity of the learning algorithms, reveals the need for important technological disruptions for the detection or reduction of the discrimination risk, and for addressing the right to obtain an explanation of the automatic decision. Since trust of the developers and above all of the users (citizens, litigants, customers) is essential, algorithms exploiting personal data must be deployed in a strict ethical framework. In conclusion, to answer this need, we list some ways of controls to be developed: institutional control, ethical charter, external audit attached to the issue of a label.

Key-words

Artificial intelligence, ethics, machine learning, discrimination impact assessment, right to explanation, quality of an automatic decision.

1 Introduction

In 2017, the CNIL launched a national debate on the theme “Numerical ethics: algorithms in debate," which led to the publication of a report. This article resumes and develops Section 4 of one of the report’s contributors’ work, Besse et al. (2017), with the goal of advancing the debate concerning the *fairness of algorithmic decisions.* A government-mandated commission chaired by Cédric Villani also published a report with the objective of "Giving meaning to Artificial Intelligence" (AI). Like the CNIL report, Villani’s commission's report focuses its attention on the ethical issues raised by the widespread, daily use of Artificial Intelligence algorithms. France is obviously not alone in its mobilization on this issue and the international initiatives are numerous, including that of the British government, which has published an ethical data framework.

Here, ours is not a question of approaching the ensemble of algorithms within AI’s vast, pluri-disciplinary fields, but rather to concentrate on those that drive decisions that impact people on a daily basis. Here, we are concerned with decisions involving banking access, insurance, health, em-

[1] Université de Toulouse INSA, Institut de Mathématiques UMR CNRS 5219.

[2] Université Toulouse Capitole, Institut de Recherche en Droit Européen, International et Comparé (IRDEIC). Membre de l’Institut Universitaire de France (IUF).

[3] Ecole Normale Supérieure de Lyon, UMPA UMR CNRS 5569.

[4] Université de Toulouse Paul Sabatier, Institut de Mathématiques UMR CNRS 5219.

ployment, and judicial and law enforcement applications. More precisely, this concerns machine learning algorithms that are trained on data sets to minimize certain statistical criteria, such as an average error rate, in order to automate decisions.

Schematically, the ethical questions mainly concern problems of data confidentiality that revolve around the learning set, the hindrance of competition, decision transparency/explainability, and their risks of discriminatory bias towards sensitive groups and individuals.

With the introduction of the GDPR (General European Data Protection Regulations n°2016/679/EU), the CNIL is concentrating its actions on the core of the issues within the field, (that is, on the protection of individual data) by proposing measurement tools to affected companies in order to evaluate incurred data confidentiality risks under what is known as the DPIA, or the *data protection impact assessment.* Indeed, it is up to the companies to be proactive on this matter and to be able to demonstrate, in the event of inspection, that they have full control over personal data security throughout the processing chain, from data acquisition to the final decision. Should any violations be found, the resulting consequence will be very severe financial sanctions: up to 20 million euros, or, for companies, 4% of global annual turnover— the highest of the two figures will be applied.

Following the adoption of Law No. 1321-2016 for a Digital Republic, which takes into account several provisions of the GDPR, INRIA has proposed a collaborative platform project (TransAlgo) that would allow for archiving automatic tools produced by five working groups:

1. Information filing engines and referral systems;
2. Learning: cracking down on data and algorithmic biases, reproducibility, explanation and intelligibility;
3. Data protection and control of data usage;
4. Metrology of communication networks;
5. Influence, misinformation, impersonification (photos, voice, and conversational agents), nudging, and fact-checking.

In this article, we propose elements of analysis, as well as tools to develop point 2 of the discrimination risks, as well as explainability, repeatability, and the quality of algorithmic and automatic decisions.

- *Discrimination*: The law protects individuals from discriminatory practices, but how can it protect them from algorithms? While the law does not mention group discrimination, the Villani report calls for the creation of a DIA (discrimination impact assessment) in section five, without referring to the already significant number of American studies around the subject. We therefore ask which tools are available in the event of group discrimination.
- *Explainability*: A careful analysis of legal texts shows that there is currently relatively little legal constraint when it comes to algorithmic transparency. Nevertheless, accepting the use of AI and automatic decisions that impact people imperatively requires elements of transparency. In this case, it concerns the right to explanation of algorithmic decisions. What could this entail?
- *Quality*: As is the case in the GDPR, French law never refers to notions of quality or the error risks associated with automatic decision-making. When it concerns the publication of results of public opinion surveys in France, French law requires that the conditions of the survey (sample size, margin of error, etc.) be made known. When dealing with AI it would also

be relevant for the law to require that the user be informed of risks associated with the application of automatic machine learning decisions. In which context should this be applied?

As the Villani report reminds us, ethical discussions have entered into the gap between what AI is capable of, and what is permitted by law. Villani highlights that, "It takes much more time to generate law and norms than it does to generate code." What's more, while the notion of a platform's fairness is present in the Law for a Digital Republic, the principals of algorithmic fairness have become ethical and legal issues in the absence of more specific legislation. This is not a product of commercial companies' altruism, but rather a necessary step in developing the indispensable trust of the larger public regarding the use of these technologies. To understand some of the problems raised by this technology, one can examine the spread of the Linky smart electricity meters, the issues related to the implementation of ParcoursSup in France (algorithmic higher-education admissions software), or even Facebook's stock downturn following the Cambridge Analytica case.

Firstly, it is important to better define how these ethical notions can be translated into technical terms. Below, section two goes on to describe in detail statistical learning algorithms, which is one of the branches of AI explored in the article. Section 3 describes the legal context and the means available to individuals and groups to protect themselves by basing the definition of disparate discrimination impact on that of the literature and its recent developments, particularly when it concerns correcting learning biases. Section 4 deals with the right to explanation in regard to the technical capabilities of commonly used statistical models and learning algorithms. Section 5 reflects on how risks of error are estimated and minimized; for example, the consequences of a 30% error are not the same when evaluating the risk of recidivism of a prisoner versus evaluating the possible interests of a given Netflix user. Finally, after attempting to summarize this complex situation, we conclude by discussing several possibilities around institutional monitoring, self-monitoring (ethical charter), and external monitoring (audits) involving the issuance of quality labels.

2 Which AI? Which Algorithms?

Artificial intelligence covers a vast disciplinary field and concerns numerous types of algorithms. We are particularly interested in those which are commonly used in our daily lives and which can lead to high-impact decisions based on personal data. Here, we discuss machine learning, which involves decision-making based on a record of known or observed situations taken from a sample that may be of varying volume. These data samples are known as *training datasets*. Amongst these, we have chosen to leave aside questions on reinforcement learning algorithms (e.g. AlphaGo) and sequential algorithms, whose applications commonly involve online commerce (bandit algorithms) and entail less serious consequences. Instead, we shall concentrate on a subset of machine learning known as *statistical learning*.

2.1 Usage Examples

Whether it be a choice of medical treatment, commercial action, preventative maintenance action, the approval or rejection of a loan, or the decision to monitor a particular individual, all of the resulting decisions are the consequence of a prediction. Examples abound in our daily lives: determining the probability of a diagnosis, the risk of a breach of contract by a client (churn/attrition rates), predicting the failure of a mechanical system, determining the risk of a client defaulting on pay-

ment, or even the risk of an individual's political or religious radicalization. These risk predictions, otherwise called scores (e.g., credit scores) are produced by *statistical learning algorithms,* which generate predictions after undergoing dataset training.

2.2 Statistical Learning Algorithms

In the 1930s, and notably following the work of Ronald Fisher, statistics was developed with a primarily *explanatory* purpose and aimed to assist decision-making. Consider the following examples: testing the effectiveness of a molecule and therefore a medication, comparing the yield of seeds in order to better choose a fertilizer, or demonstrating the influence of a given factor (tobacco, sugar consumption) on public health objectives. Decision-making is thus the consequence of a *statistical test* that measures incurred *error risk.* But it just so happens that these same statistical models can also be used for purposes that are merely predictive: predicting the concentration of ozone in days to come, predicting the risk of a company's payment default, etc. Moreover, these statistical models can be sufficiently simple (typically linear), to be easily *interpretable.*

Nevertheless, certain situations and phenomena require more complex models, commonly known as algorithms, if they are to be correctly approached in order to produce sufficiently reliable predictions. Since the end of the 1990s, all scientific disciplines from statistics to mathematics and computer science have contributed to the development of a vast array of various algorithms with an essentially *predictive aim.* Among these were binary decision trees, *k* nearest neighbors, vector support machines, neural networks, and eventually deep learning, random decision forests, and gradient booting machines, to name a few. It is no longer a question of testing the influence of a factor or the effectiveness of a treatment; here, what matters is the quality of the prediction. The literature on this subject is vast: consult, for example James et al. (2017), or the pedagogical resources on the website wikistat.fr.

2.3 The principle of statistical learning

The principle of statistical learning algorithms is based on the fact that, from a set of examples called a training dataset, it is possible to develop a decision-making rule that will apply to all future cases. From a large quantity of collected data that mainly contains decisions that have already been made and the variables that explain them, mathematical principles allow not only to understand how the decisions were made, but also to remove the rules that directed them.

In concrete terms, identifying these rules consists of finding trends (patterns or features) within the observations. It is necessary to detect the behavioral characteristics in the data that make it possible to segment individuals into homogeneous groups. So, depending on our characteristics, our profile type can be defined in relation to other previously analyzed individuals, and the algorithm will then emit a fixed rule according to the group of membership and according to any similarities or resemblances to previously analyzed individuals. The process of identifying standard behaviors is automated, and there is no monitoring after the fact. It is from these standard behaviors that models are created, decisions are made, and future events are predicted.

More precisely, model parameters are estimated and algorithms are *trained* and optimized on training data sets in order to minimize any prediction or generalization errors. This error rate is normally estimated by the calculation of a statistical average of errors made, known as the *mean* or *average error rate,* on test data sets independent of the training dataset. A statistical learning algo-

rithm best adapts itself to historic data so that it may identify the specificities of current or pending data. It then produces the most suitable prediction, without possibility of creativity; schematically speaking, in order to make the most accurate prediction, this involves identifying the past situation which most resembles the current situation.

2.4 Risks of statistical learning

It is important to note that the training data set has to be all the more voluminous, as the algorithm's complexity lies both in the number of parameters to estimate, as well as the parameters that define it. Consequently, the enormous increase in computing and archiving capabilities, combined with the explosion of available data volume has led to very significant advances in the quality of learning algorithms and the decisions that result from them. One of these advances is in that of deep learning, which has seen enormous development since 2012. In this sense, the current success and the media hype around statistical learning, and AI in general, are a direct consequence of the *datafication* of our daily lives, which seeks to use the data from our messages, web-engine searches, purchases, travel and more, all in the goal of systematically storing it.

This announces the advent of the Big Data era paradigm. In traditional Cartesian reasoning, a theory makes it possible to produce a model originating from human thought. Then, the model is put into action and presented with data collected from experiments that were performed specifically to confront the model with data.

Thus, the theory can be clearly refuted or accepted using fact as a basis. The model can then be analyzed from a moral or ethical point of view, and perhaps be discussed. But in learning, the creation of a model originates from a database study, without *a posteriori* analysis. We thus understand that from the moment that we decide to entrust an algorithm with decision-making power, it can shape reality to conform to its model. It freezes reality, so to speak, according to what it has seen through the prism of the samples provided during the learning phase, and subsequently reproduces the model infinitely. Naturally, the model no longer evolves and comes to adjust reality so it matches its own predictions. Once behavior is learned, the rules of prediction can be expressed thus: no longer is anything left up to chance or creativity; here, repeatability reigns.

Often, being confronted with new ideas allows each person to clarify their own Truth, all while becoming aware one's own errors— even when one ultimately consciously makes a wrong choice. But AI is categorically different: the algorithmic matrix seeks to optimize decisions, "justly or coldly". Naturally, the morality and equity behind the judgement is not predefined, but depends firstly on the manner in which rules are learned (the chosen objective criterion), and secondly on the manner in which the learning sample was created. The choice of mathematical rules allowing for the creation of the model is essential.

A rather delicate question then arises: How, or by which "measurable" characteristics can we translate notions of fairness, trustworthiness, and accountability to such algorithmic decisions when they are the consequence or result of a prediction?

The answer can be broken down into three points:

- The decision must avoid all discriminatory bias towards minorities and vulnerable, legally protected groups
- Whether the decision be statistic or probabilistic, it must be possible to attribute this decision to a human who *assumes responsibility* for it. He or she must be able to account for it, and be able explain it, as a doctor does to his patient.
- The decision must be as accurate as possible, in the interest of the person and/or the community concerned, thus resulting in a better decision.

Let us consider these three points in detail.

3 Bias and discrimination

Part 5 of the Villani Report, devoted to ethical questions, emphasizes the risks of discriminatory algorithmic practices that reproduce or even reinforce societal biases. This section will focus on notions of individual and collective discrimination, in order to lay a foundation for:

- the measurement of discriminatory biases in order to
- build tools to detect biases,
- and build tools to possibly correct them.

3.1 Juridical Framework

According to article 225-1 of French penal code, "Discrimination comprises any distinction applied between natural persons by reason of their origin, sex, family situation, pregnancy, physical appearance, particular vulnerability resulting from their visible or known economic situation, surname, residency, state of health, loss of autonomy, handicap, genetic characteristics, morals, sexual orientation, gender identity, age, political opinions, union activities, capacity to express oneself in another language than French, true or supposed membership or non-membership to a given ethnic group, nation, presumed race or religion."

Article 225-2 adds that, "Discrimination defined by article 225-1, committed against a natural or legal person, is punished by three years' imprisonment and a fine of €45,000, especially where it consists:

1. of the refusal to supply good or service;
2. of obstructing the normal exercise of any given economic activity;
3. of the refusal to hire, to sanction or to dismiss a person."

French law only addresses an *individual approach* of the notion of discrimination risk. The Villani report asks that we consider *group* discrimination and stresses the necessity of defining an evaluation tool. The report references the Discrimination Impact Assessment (DIA), a compliment of the Data Protection Impact Assessment (DPIA), drafted by the GDPR, which protects personal data of individuals and not of groups. While this is not addressed in the Villani report, there is abundant literature on this subject, especially on the subject of *disparate impact,* which has been studied in the US since the 1970s.

For its part, European framework strictly regulates the collection of sensitive personal data (religious or political affiliation, sexual orientation, ethnic origin, etc.) and forbids those responsible for algorithmic decisions to use them in automatized processes (art. 22§4) without the explicit content of the person, or without substantial public interest. A decision is declared fair if it is neither based on affiliation to a protected minority group, nor based on the explicit or implicit knowledge of sensitive personal data.

This point is without a doubt the most difficult to clarify. Indeed, even if the "sensitive" variable is unknown, or even deleted from the training data set, the decision is not necessarily without bias. Sensitive information can be contained implicitly, even without the intention to look for it. It can be hidden in non-sensitive data, and thus participate in bias in the decision-making process. Consumer habits, opinions on social networks, and geolocation data all provide information about a person's beliefs and identity, and can implicitly become sensitive data.

The questions raised and difficulties encountered during algorithmic construction with a goal of fairness are directly related to the training conditions involving decision-making. In effect, as noted in section 2, any learning that occurs is a reflection of the training data base.

Consequently:

- If the data itself is biased and not representative of the population, or
- if a structural bias remains present in the population,

then this is the source of a breach of equity. The decision reproduces and can even reinforce the bias, thus leading to discrimination. More dangerous still, the decision becomes a self-fulfilling prophecy. If too high, an estimate of credit risk generates a higher rate, hence higher repayments, which increase the risk of default. An inflated risk of recidivism delays release, increases desocialization, and ultimately reinforces the risk of re-offending. Cathy O'Neil (2015) develops on the perversity of the unintended consequences of these types of tools. It's worth noting that classification algorithms seek to separate populations into sub-groups. Thus, if this separation is already present in the data, the algorithm will learn and amplify this dissimilitude, consequently introducing treatment discrimination into the data.

Figure 1 illustrates an example of severe bias in a pan-genome database (genome-wide association study, GWA study, or GWAS). These databases archive analyses of genetic variations (singular nucleotide polymorphisms or SNPs) in a significant number of individuals, in order to study their correlations with phenotypic traits, for example illnesses. These studies laid the primary foundation for research on personalized therapies.

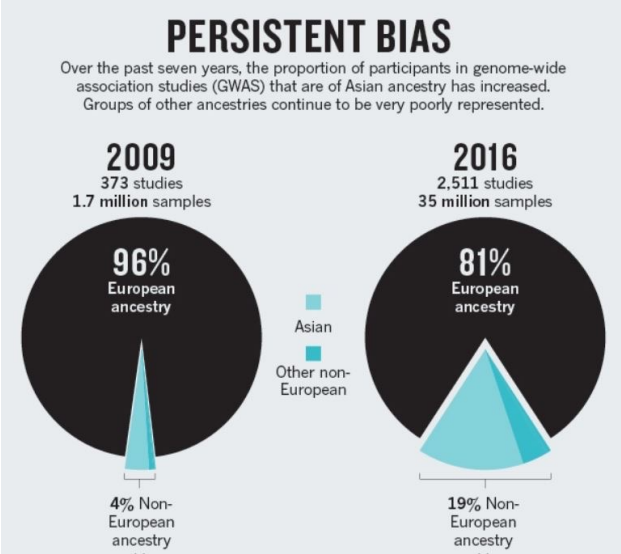

Figure1. Panpejoy and Fullerton (2016): Bias in pa-genome (GWAS) databases.

From an ethical point of view, the problem is that the great majority of GWAS databases were developed on populations of Caucasian/European descent (Figure 1, Popejoy et Fullerton, 2016). The risk factors estimated by classic statistical models (logistic regression) or by automatic learning algorithms would likely be much less precise for a patient of African or Asian descent. In other words, at this point in time, African and Asian descendants can expect very little from personalized treatments based on genomes.

3.2 Individual discrimination

Proving individual discrimination is particularly difficult for the concerned person, unless they accept the use of probationary procedures that are, in principle, prohibited because they are considered unfair. This may be done with the help of testing devices used in situation tests and discrimination tests, in order to detect instances of discrimination. Such devices serve both as a means of investigation and a form of social experimentation in real life situations, and may include sensitive data such as ethnic origin, disabilities, sex, sexual orientation, religion, and union membership. This type of test does not respect the principle of the fairness of proof, but it is the most efficient, and often the only way, to prove discrimination. In the simplest of cases, it's a question of comparing the behavior of a third party towards two people with exactly the same profile and all of the same pertinent characteristics, with the exception of those characteristics that one would imagine could lead to discrimination. Naturally, when discrimination is not based on one single or even several sensitive data, but is instead the result of cross-referenced data that can indirectly allow discrimination, we must understand the results in all of their complexity.

This method, used by associations such as SOS Racism, is recognized by French courts in the measure that, although considered to be an unfair practice, it cannot be dismissed as a means of seeking proof, as declared by a judgment of France's Court of Appeals in June 2002 in the *Pym's de Tours* case. In the judgement, the solution referred to Article 225-3-1 of the Penal Code, according to which: "The offenses provided for in this section are established even if they are committed against one or more persons having solicited any of the goods, acts, services or contracts referred to

in Article 225-2 for the purpose of demonstrating the existence of discriminatory conduct, provided that proof of that conduct is established".

The principle is therefore quite simple. For example, in the case of a hiring procedure, it suffices to send two unique CVs at different times to respond to job offers. The CVs, which may indicate the origin, name, or age of potential applicants should not be different except for the name. In order to be valid, the CV and the application must be genuine; consequently, only the competing CV is modified. Several organizations and laboratories practice these types of operations: take, for example, the *Observatoire des Discriminations*, the TEPP of University Marne la Vallée (L'Horty et al. 2017) or even DARES, The Directorate for Research, Studies and Statistics (*Direction de l'animation de la recherche, des études et des statistiques*) under the Ministry of Work, in association with the group ISM Corum. Certain companies even ask ISM Corum to test their recruitment processes.

In another vein, Galhotra et al. (2017) define the bias of a decision by including a notion of causality. The relative software (Themis) runs an automatic test for discrimination resulting from software. This process may be replicated or simplified to be applied to algorithmic decisions in order to evaluate the risks of discrimination against individuals. It merely requires running a test sample a second time, this time switching the two categories of the sensitive variable (for example gender or ethnic origin). The algorithm is then applied once again to the modified test sample in order to identify the individuals for whom the decision has been changed and, for example, become more favorable, with the simple change of gender, age, or ethnic origin. Even if the number of changes found is low and not statistically significant, these persons were clearly discriminated against; the algorithm is the source of unfair discrimination due to possible bias in the sample. Situations like this could risk costly legal disputes for companies that use such algorithms.

Make no mistake; the sensitive variable must be known. Obviously, this is not always the case, and presents a problem, as removing the model's sensitive variable does not necessarily prevent a discriminatory decision, but prevents being able to simply identify the bias. The second difficulty is that it is also a question of providing evidence of *discriminatory intention*. However, in the case of discrimination by algorithmic processing, discrimination is not necessarily the result of intent.

3.3 Group discrimination

The Villani report recommends the creation or definition of a discrimination measure on a group level, based on the notions of the Discrimination Impact Assessment (DIA), and not only a definition in the legal, individual sense. While the academic literature on the subject proposes many manners in which to measure positive or negative bias towards a person's membership or non-membership to a group (generally this is a minority group protected by law), the first difficulty here lies in the choice of a measurement of discrimination. An individual measurement type favors similarity in the sense of k-nearest neighbors of an individual, in order to detect an atypical situation. Nevertheless, this individual may be surrounded by those who belong to the same protected group, and not all would wrongly benefit from a positive decision. It is more informative to consider a collective or statistical measure of discrimination based on a contingency table (see table 1) which

crosses two variables: the sensitive variable (membership in a group protected by law) and the reception of a Positive (credit, job, stock market) or Negative decision.

Table 1. Contingency table between group membership and outcome of a decision. Associated proportions: p1=a/n1, p2=c/n2, p=m1/n

Protected Group	**Decision**		
	Positive	**Negative**	**Margin**
Yes	*a*	*b*	*n1=a+b*
No	*c*	*d*	*n2=c+d*
Margin	*m1*	*m2*	*n=n1+n2*

Simple measures of discrimination are defined with this table (Pedreschi et al. 2012):

- Risk difference: *RD = p1-p2*,
- Relative Risk: *RR = p1/p2*,
- Relative Chance: *RC = (1-p1) / (1-p2)*,
- Odds Ratio: *RR / RC*.

But many other measures are proposed. Consult, for example, Žliobaité (2015), who addresses differences of means, regression coefficients, ranking tests, mutual information, and prediction comparisons. The problem is that there are far too many possible technical and statistical definitions of discrimination, but at this point in time, there is no known legal base to justify the choice of any one definition.

3.4 Disparate impact

We shall next focus on the common, widespread definition of disparate impact, which is defined as the ratio of two probabilities:

$$DI = \frac{P(Y=1|S=0)}{P(Y=1|S=1)}.$$

It is the ratio of the probability that a decision will be positive *(Y=1),* knowing that the group is protected *(S=0)* to the probability that the decision will be positive (*Y=1),* knowing that this other group is not protected *(S=1).* It is estimated by relative risk (*RR*), defined with the help of the contingency table (table 1 above).

Feldman et al. (2015) provide several historical elements[5] of the first use of this criterion by the court system of the state of California, dating back to 1971. Its use has inspired many articles in law reviews, notably regarding operation mode, which consists of comparing obtained value to an empirically fixed ceiling at 0.8. Above 0.8, the impact is judged sufficiently disproportionate and is an indicator of discrimination. The same reasoning is used pursuant to Title VII of the Civil Rights Act of 1964, which is a federal law that prohibits employers from discriminating against employees

[5] Consult the relative page on Wikipedia.

on the basis of sex, race, color, national origin, and religion. Nevertheless, if the company provides proof that their recruitment choices are based on criteria that are necessary to the "economic interests" of the company, the discrimination is not considered to be illegal.

In summary, in the US, a DI evaluation makes it possible to reveal situations that are overly disproportionate and are to the detriment of a sensitive or protected group. This opens up the possibility for detecting or even sentencing a company for implicit group discrimination. In France, and even in the larger European community, the legislation only recognizes individual discrimination and has very rarely recognized collective discrimination.

The DI evaluation raises another statistical question, as emphasized by Peresie (2009). Is it necessary to compare the equality of DI terms with a statistical test, in order to introduce incertitude or simply compare the DI to the ceiling of 0.8? These two strategies can lead to contradictory results. What's more, the equality test is based on a hypothesis of normality which is ill-advised in itself. To avoid these difficulties, Besse et al. (2018) propose an estimation of the DI by confidence interval, including a statistical measurement for error risk, without the use of a normality hypothesis. The exact asymptomatic distribution is obtained with the application of the central limit theorem and with the linearization of the fairness criterion.

Unfortunately, the evaluation and characterization of discrimination against a group cannot be limited to a simple DI evaluation. Thus, facial recognition algorithms are regularly accused of racism, but on the basis of another criterion: that of the error of recognition. Presumably this is because there are learning bases in which some groups, especially women of African descent, are largely underrepresented. Consequently, error rates can reach up to 30%, as opposed to 1% for a man of European descent.

Furthermore, even if the DI is limited and the error rates for all sensitive variable categories are identical, another source of discrimination can spread in the dissymmetry of an algorithm's or predictor's confusion matrices. This viewpoint is the basis of the *What-If Tool* used by Google which is available on their platform. The adopted discrimination measurement is thus the *equality of opportunity* of a learning algorithm, as described by Hardt et al. (2016). Besse et al. (2018) also offer an estimation of discrimination measurements by confidence interval, referred to as *Conditional Procedure Accuracy Equality*.

3.5 Example: COMPAS Risk of recidivism

This approach is well-illustrated by the controversy between the website *ProPublica* (Pulitzer Prize 2011) and the company Equivant, formerly known as *Northpointe.* This company, taking an approach of "predictive justice" commercialized the application *COMPAS* (Correctional Offender Management Profile for Alternative Sanction), which produces a score, or risk of recidivism for detainees and defendants during a trial. ProPublica has accused this score of being biased, and therefore racist. This controversy has given rise to a great number of articles, all of which serve to reinforce a large volume of work that has existed around the subject for some twenty years. These studies highlight prohibitive contradictions between the proposed criteria. Let us now summarize the controversy.

The recidivism score is estimated on the basis of a detailed questionnaire and from a life expectancy model known as the Cox model. The quality of this score is optimized and measured by

the AUC coefficient (an area under the ROC curve), approximately around 0.7, which is a rather weak value in relation to the high error rates that have been observed (30-40%). The company *Northpointe* defends the fairness of the score, assuring that:

- The distribution of its values (and therefore the selection rates) is comparable according to the origin of the accused (African-American vs. Caucasian); the DI is therefore insignificant.
- The resulting recidivism prediction error rate (the confusion matrix) is similar between data subjects according to their origin, around 30 to 40%; the argument used for facial recognition is not admissible.

For their part, Angwin et a. (2016), from the site *ProPublica,* have denounced a bias in the COMPAS scores by studying a group of freed detainees for which the COMPAS recidivism score was known, and by observing if there were any instances of arrest over a two-year period. They showed that the rate of *false positives* (elevated *COMPAS* score, without observed recidivism) is *much higher* for former prisoners of African American origin than for those of Caucasian origin. As the COMPAS score can potentially be used to set terms of parole and bail, a detainee of African American origin is more likely to stay in prison longer, risking the reinforcement of their desocialization, and thus increasing their risk of recidivism.

To explain the stalemate faced in this controversy, Chouldechova (2017) shows that under the constraints of "fairness" as monitored by *Northepointe,* and, considering that the rate of recidivism of African Americans is indeed higher, the false positives and negatives can only be considered unequal and to the detriment of African Americans. This becomes all the more evident, as the error rate (40%) is significant.

The question that arises (or that should have already been asked) concerns the quality of this prediction. Under the apparent objectivity of an algorithm, there lies a significant error rate that largely discredits the COMPAS tool. Using a group of individuals lacking judicial expertise, Dressel and Farid (2018) demonstrated in a web interview that the COMPAS predictions were equally unreliable, just as a simple linear model involving only two variables would be.

3.6 Data repair to foster fair algorithms

To detect unfairness in algorithms, we have seen that it is possible to compute numerous criteria, each highlighting a type of differential treatment between diverse subgroups of the population. This detection can be displayed as an imbalance between the proportion of accurate predictions within each subgroup and of a difference of error distribution or other criteria which demonstrate a dependent relationship between the learned decision and the sensitive variable which divides the population into two subgroups. Thus, the notion of total fairness must be characterized by independence between these two laws of probability (del Barrio et al. 2018b). The stronger the link is, the more pronounced the discriminating effect will be. This formalism has led various authors to offer several ways to remedy this breach of equity, either by changing the decision rule, or by changing the learning sample. Modifying the rule amounts to preventing the algorithm from over-learning this link, by imposing a term that favors the absence of a link between the prediction and the sensitive variable. (Zafar et al. 2017). Modifying the sample means fostering independence between the data and the

sensitive variable to ensure that any algorithm that uses these data as a learning base cannot reproduce bias against the sensitive variable.

To be able to do this, it is necessary to modify the conditional laws in relation to the sensitive variable, and to make them as similar as possible, without losing too much information and possibly harming the model's predictive capability. This solution, described in Feldman et al. (2015) was studied by del Barrio et al. (2018a). Nevertheless, there is a price to pay to achieve non-discrimination: that is, constructing a rule that is less predictive with respect to the learning sample. The statistician must therefore control both the error made by the prediction rule, as well as the desired non-discrimination.

4 Explainability of a decision

4.1 Laws and challenges of the Right to Explanation

The Villani report called for "opening the black boxes" of AI, as a large number of the ethical questions raised revolve around the opacity of these technologies. Given their growing, not to say invasive, position, the report considers that this is a democratic issue. Article 10 of Law No. 78-17 On Information Technology, Data Files, and Civil Liberites of January 6th, 1978, originally provided that, "No other decision having a legal effect on an individual may be taken solely on the grounds of automatic processing of data intended to define the profile of the data subject or to assess some aspects of their personality." In other words, an automated assessment of a person's characteristics leading to decision-making cannot be achieved merely on the basis of automated processing. This therefore implies that other criteria will be taken into account, and even that other methods will be used. In particular, those affected by the decision can expect that the evaluation can indeed be verified through human intervention. Though this principle which tends to control the negative effects of profiling has long been established, its explanation was not enough to prevent the explosion of this technique, concurrently with the emergence of the massive collection of data on the Internet. Many profiling techniques have been developed without necessarily providing technical or human safeguards. In consequence, this rule is poorly respected, and for the moment, its violation has not led to sanctioning.

Similarly, the GDPR and directive 95/46/CE which preceded it establish a number of rights in the event that individual decisions be made on the basis of automated processing:

1. The right of access and the right to be informed of the existence of automated decision-making (GDPR, art. 13-15);
2. 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" (GDPR, art. 22§1);
3. "The right to obtain human intervention on the part of the controller" (GDPR, art. 22§3);
4. "The right to express his or her point of view and to contest the decision" (GDPR, art. 22§3);

In principle, any sensitive data must be excluded from exclusively automated processing (art. 22§4), except when explicit consent is given, or if there is a question of public interest.

However, several exceptions (GDPR, art. 22§3) have been planned for, if the decision:

a) "Is necessary for entering into, or performance of, a contract between the data subject and a data controller;

b) Is authorized by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

c) Is based on the data subject's explicit consent."

This series of exceptions is far from trivial, and substantially weakens the rule. Indeed, with regard to digital economic activities, numerous automated processes can claim a contractual basis, as an Internet user's interaction with services on e-commerce sites and linking platforms (such as social networks) is considered as an acceptance of general conditions of use, and constitutes the acceptance of a contractual offer. In addition to these digital activities, one can consider the previously cited hypothesis regarding access to credit, housing, goods and services, all of which are most often based on a conclusion of contract.

Furthermore, point c) of the previous paragraph provides for the assumption of the data subject's explicit consent. Though consent in itself can be rather easily acquired, we can nevertheless doubt that it is *informed* consent, as intellectual accessibility to automated processing methods is unlikely for the laypersons that make up the vast majority of the data subjects, particularly when their consent is acquired online.

These provisions have been incorporated into French law with the recent adoption of law number 2018-493 of June 20th, 2018, which amended law number 78-17, known as "*Informatique et Libertés*" of January 6th, 1978. Article 21 amends Article 10 of the law of January 6th, 1978, in order to extend the cases in which, exceptionally, any decision leading to legal effects or any decision having a significant effect on a person may be taken on the sole basis of automatically processed personal data. Article 10, paragraph 1 of law 78-17 henceforth provides that, "No legal decision involving an assessment of a person's behavior may be based on automated processing of personal data intended to evaluate certain aspects of that person's personality."

Paragraph 2 adds that, "No decision which has legal effects on or which significantly effects a person can be taken solely on the basis of automated processing of personal data, including profiling." Two exceptions to this principle are provided for. The first refers to the exceptions to the GDPR, that is to say, "the cases mentioned in sections 2a and 2c of the aforementioned Article 22 are subject to the reservations mentioned in paragraph 3 of this same article, provided that the rules defining the processing, as well as the principal characteristics of its implementation are communicated, with the exception of secrets protected by law, by the person responsible for the data processing to the data subject if he or she so requests".

In addition to the guarantees provided by the European text in Article 22§3 (the right to obtain human intervention from the controller, right to express one's point of view and to challenge the decision), French lawmakers have added the obligation to communicate the rules defining the data processing, as well as the main characteristics of its implementation at the request of the person concerned. This guarantee ceases to be valid if these rules are subject to legal confidentiality. This

reservation here also substantially weakens the principle, though communication of the rules that preserve legal confidentiality could easily be envisaged.

As for the second exception provided for in Article 10, paragraph 2 of the amended Law 78-17, it emphasizes point B of Article 22§2 of the GDPR, under which every member state may freely provide exceptions, as long as they are legally provided for and respect certain guarantees. French lawmakers made an exemption for individual administrative decisions, under the condition that the processing not be based on sensitive data, that administrative recourse be possible, and that the information be provided with the use of an algorithm. This particular exemption was already established by Article 4 of Law 2016-1321 for a Digital Republic, codified in Article L. 311-3-1 of the CRPA, according to which an individual administrative decision made on the basis of algorithmic processing must include an explicit mention which informs the concerned party. Article 1 of Decree No. 2017-330 of 14 March 2017, codified in Article R. 311-3-1-1 CRPA, specifies that the explicit mention must indicate the pursued aim of the algorithmic processing. It specifies the right to obtain communication of the rules defining the processing and the main characteristics of its implementation, as well as the procedures for exercising this right to communicate and the right to referral, where appropriate, to the Commission for Access to Administrative Documents. Law 2018-493 of June 20th, 2018, came to specify that the aforementioned explicit mention is required under penalty of nullity. The sanction of the violation of this obligation of information is thus explicitly foreseen.

Since the adoption of the Law for a Digital Republic of October 7th, 2016, Article L. 311-3-1 additionally provides that, "The rules defining the processing as well as the principal characteristics of its implementation be communicated by the administration to the interested party, should he or she request so." Decree number 2017-330, codified in Article R. 311-3-1-2, specifies that the information be provided in an intelligible manner, without infringing on legally confidential information. 1st: The degree and mode of contribution of algorithmic processing to decision making; 2nd: The data processed and their sources; 3rd: the treatment parameters, and if applicable, their weighting in regards to the situation of the interested party; 4th: The operations carried out by the processing.

Law 2018-493 goes further in regard to the use of an automated processing system for administrative decision-making, and now provides for the requirement of explanation. It provides that, "the person responsible for data processing ensures complete control over the automated algorithmic processing and its evolution in order to explain intelligibly and in detail the manner in which the processing results will be implemented in respect to the person concerned." The famous "right to explanation" is explicitly established in French law, whereas the GDPR only makes a clear reference to it in recital 71. Articles 13 to 15 merely provide for the right to information and access on the use of an automated device and its "underlying logic", which constitutes a very general approach, and is disconnected from the individual situations of the persons concerned.

Notwithstanding this exception in favor of the administration, no decision the administration decides on an administrative appeal can be made on the sole basis of automated processing of personal data.

Law number 2018-493 was the subject of a decision of the Constitutional Council No. 2018-765 DC on June 12, 2018, notably on the aspects concerning the automated individual decisions

made by the administration (points 66 et seq.). The Constitutional Council considers that the provisions of the law are limited to authorizing the administration to proceed in the individual assessment of the citizen's situation by the mere invention of an algorithm, according to the rules and criteria defined in advance by the person responsible for processing. They have neither the goal nor the effect of authorizing the administration to adopt decisions without legal base, nor to apply any rules other than those of existing law. As a result, there is no loss of regulatory jurisdiction power (point 69).

Secondly, the mere use of an algorithm as the basis of an individual administrative decision is subject to the fulfillment several conditions: On one hand, the individual administrative decision must explicitly mention that it was made on the basis of an algorithm and the principal characteristics of its implementation must be communicated to the interested party upon his or her request. As a result, when an algorithm's operating principles cannot be communicated without violating confidential legal information, no individual decision can be taken on the sole basis of this algorithm (point 70).

Furthermore, it must be possible to subject the individual administrative decision to administrative repeal. The administration called upon in the event of appeal is required to make a ruling without exclusively basing their decision on the algorithm. In the event of litigation, the administrative decision is placed under the control of the judge, who is likely to require the administration to communicate the algorithm's characteristics.

Finally, the exclusive use of an algorithm is forbidden if the processing is based on one of the sensitive data mentioned in paragraph I of article 8 of the law of January 6th, 1978, that is to say *personal data,* "which reveal alleged racial or ethnic origin, political opinions, religious or philosophical convictions, or the union membership of a natural person", genetic data, biometric data, health data, or data relative to the lifestyle or sexual orientation of a natural person (pt 70).

Finally, the person responsible for processing must ensure full mastery of the algorithmic processing and its evolution, in order to explain to the concerned person— in detail and in an intelligible form— how the data processing was implemented. It follows that algorithms which are susceptible to revising the rules that they themselves apply cannot be used as the basis of an individual administrative decision without the control and validation of the person responsible for processing (pt 71). The Constitutional Council considers that the lawmakers have defined appropriate guarantees for safeguarding the rights and freedoms of individuals subject to individual administrative decisions made on the sole basis of an algorithm. Article 10 (2) of the Law of January 6th, 1978 is in conformity with the Constitution (pt 72).

Schematically speaking, different situations can be considered for the application of these rules. In the case of a procedural algorithm such as *PacoursSup*, the rules of operation must be clearly explained. The concerned Ministry prepared for this following difficulties encountered by the predecessor algorithm APB. Indeed, the code of the *ParcoursSup* algorithm is certainly available to the public, but the surrounding debate and controversy remain, as the rules that determine individual establishments' decisions can remain confidential, thus making the process opaque and possibly discriminatory. Finally, law number 2018-493 provides, particularly when it comes to decisions made in regard to the education sector within the framework of ParcoursSup, that "the ethical and scientific committee referred to in article L.612-3 of the education code annually submits a

report to Parliament at the end of the national pre-registration procedure before December 1st, explaining how the procedure is conducted and the procedures that higher education institutions use when examining applications. It is on this occasion that the committee can make any proposal to improve the transparency of this procedure".

4.2 What transparency?

While the goal of the provisions of the GDPR is to strengthen the rights of the persons concerned, shortcomings can be identified. On one hand they are related to exceptions within the GDPR, and on the other hand, they are related to the fact that transparency in the wording of these rights is in not required. The only reference to a "Right to Explanation" provides that the data subject has the right to obtain information from the processing manager that confirms the existence of automized decision-making, including profiling, referred to article 22, paragraphs 1 and 4, but also, "at least in such cases, useful information concerning the underlying logic, as well as the significance and the expected consequences of this processing on the data subject". It can therefore be said that the general rule on data protection neither directly nor indirectly truly concerns the principal of algorithmic transparency.

The law for a Digital Republic was essentially intended to impose an obligation of information (loyalty) on referencing algorithm methods, which is added to the other information requirements in the Consumer Code. Above all, this obligation is usefully completed by the pre-existing provisions in the Consumer Code relating to deceptive marketing practices whose claims are sufficiently broad to target and sanction deviant behavior that could be founded on unfair or inaccurate algorithmic processing; the right to explanation only explicitly concerns the administration. On the other hand, law number 2018-493 of June 20th, 2018 put in place further requirements of transparency and explanation. It is still early on to know how effective these measures will be on the future of algorithmic and AI transparency, but French lawmakers are particularly ambitions in comparison to their EU counterparts and to the other EU member states.

4.3 What explanation?

In the following section, we state that an algorithmic decision is interpretable if it is possible to explicitly report on the known data and characteristics of the situation. In other words, it is possible to relate the values taken by certain variables (the characteristics) and their consequences to the forecast, for example on a score, and thus on a decision. On the other hand, an algorithmic decision is said to be explainable only if it is possible to identify or quantify the importance of the characteristics or variables that make the greatest contribution to the decision.

In the case of an opaque algorithm, it is impossible to simply relate values and characteristics with the results of the decision, notably in case of a non-linear model or a model that has many interactions. Such a high-value variable can lead to a decision in one direction or another, according to the value taken by another non-identifiable variable, or even a complex combination of other variables. An opaque model that cannot easily be explained (for example one that is used on a job applicant) leads to the decision-maker being released of all accountability, allowing him to hide behind the algorithm. Here it is no longer the fault of the computer, but of that of the algorithm.

Each actor, public or private, and each sector— health, justice, employment, banking, insurance, law enforcement— requires careful reading on what can be a form of algorithmic transparen-

cy related to the right of explanation. An *ad minima* display of ethical behavior is essential to the acceptability of these technologies, but, within this framework, the formulation of an explanation depends on many factors. Explaining an automatic diagnosis and the relative risks incurred during a surgical procedure, justifying incarceration based on an algorithm's estimation on the risk of recidivism, providing a rationale for the refusal of a loan based on a score… all of these require dual skill: skill relating to the field in question, and knowledge of the limits and properties of the algorithm that led to the decision.

4.4 Aids to explanation

What can an individual do, whether he or she be the person responsible for decision-making, or a mere citizen or client affected by it, when confronted with a set of hundreds of decision-making trees or with a network of neurons defined by thousands, or even millions of parameters learned on a large volume of data?

In many fields of application, and notably in medicine in patient-doctor interaction, an opaque model that does not allow for easy explanation and leads to the decision maker being freed of accountability would only be acceptable with great difficulty, unless it were possible to provide a significantly higher quality of prediction in the search for a better compromise between quality and explainability. In other words, it is necessary to favor an interpretable elementary model that is potentially less precise than a complex model that involves a great number of parameters and has better predictive qualities that are nevertheless opaque to all interpretation.

A graded approach could be considered, according to the priority given to the explanation of the quality of the prediction, assuming that a more opaque algorithm allows inversely for better results. The response would be potentially different according to the activities in question because it is not relevant to treat an algorithm used in medicine or in marketing techniques in the same manner. This would then lead to the encouragement of sectoral regulation. In any event, it seems crucial to be able to make a social choice on what is preferable in a balance of interests between the quality of the explanation and the quality of the forecast, at least in the case of hypotheses where the algorithmic characteristics are reducible to these main two qualities.

Let us also note that the "right to explanation" can be the subject of two different approaches (Watcher et a. 2017 p5):

- The right to have an explanation of the general functioning of the system implementing algorithmic decisions;
- the right to have an explanation of a specific decision.

Moreover, the explanation can be ex ante or ex post (Watcher et a. 2017 p6). If it is a question of giving a specific explanation of an individual decision, the explanation can only be given ex post, while if it is related to the general functioning, it can be ex ante or ex post.

In the case of an interpretable learning algorithm, the coefficients of a linear or logistic model can and must be explained to the concerned individual and to the sequence of rules that define a decision tree. The case of an opaque algorithm or a simply explainable algorithm hardly seems to be of concern or taken into account by the law.

Given the importance of the issues on explainability (opening the black boxes of AI), research in this field is highly active. Let us cite and compare several illustrative examples of algorithms in which the aim is to gain a comprehensive understanding of a complex algorithm, versus others where an individual explanation is the aim. In the first case, aids seek to identify the most important variables (features), that is to say, those which are most commonly involved in decision-making. Regardless of any ethical issue, this is fundamental to analyzing the reliability and robustness of decisions and identifying the possible artifacts that are generally the result of inadequacies in the learning base. Different strategies are proposed: For data aggregation trees (random forest, gradient boosting), it is common to look for the variables with a mean decrease accuracy that have the most degrading effect on the estimation of a decision's quality. A more general method consists of locally approximating the decisions reached by a non-interpretable algorithm to the mean of a regression-type interpretable decision rule. Indeed, for a regression or a logistic regression model, the role that each variable plays is clearly expressed by the mean of a linear combination. Each coefficient corresponds to the weight that each variable has on the prediction, and thus makes it possible to determine not only the importance of each variable, bt also if its contribution is positive or negative in the final result. A decision rule can be much more complex than a linear rule; this approximation does have meaning globally, but only locally. This methodology was developed in the LIME package (2016). A similar idea consists of testing the algorithm on an algorithm that presents bias in each variable. Thus, if we succeed in creating a nearly similar legal test sample, with the exception being that one of the variables presents a deviation from the mean (positive or negative), we can study the evolution of the decision rule in a general manner, since we consider the laws of the algorithm's output samples.This methodology answered the question of how to moderately influence an algorithm's decision by increasing or decreasing some of its characteristics. This work is detailed in Bachoc et al. (2018).

Figure 2. A husky (on the left) is confused with a wolf, because the pixels (on the right) characterizing wolves are those of the snowy background. This artifact is due to a learning base that was insufficiently representative.

This approach is still valid for even highly complex algorithms, such as deep learning, used in image recognition. An often-cited example (Tulio Ribeiro et al. 2016), highlights a weakness within the learning base. As wolves have been systematically photographed against a snowy background, it is the background that allows identifying them the most surely. A husky that was also

photographed against a snowy background is thus confused with a wolf (figure 2). The problem is different for the explanation of an individual decision. It is again a question of identifying the most representative variable or variables that, if modified, would make it possible to tip the decision. For example, this could mean identifying the revenue required to obtain a loan, or the behavioral variable that most favorably impacts the evaluation of a recidivism score. This goal could be achieved by an interpretable local approximation of the most complex algorithmic rule: approximation by a simple linear model, or by a basic binary decision tree.

In conclusion, European regulation and French law, which are mainly focused on administration, leave considerable room for maneuver with regard to transparency. This space left for maneuver remains empty, and without significant progress in the fundamental research of the subject, must be occupied by ethical actions, at the risk of provoking massive rejections of AI technologies. Thus, the *Partnership on AI,* which supports for AI in the service of humanity, created between the main industry actors (Google, Facebook, Microsoft, IBM, Apple…), is very sensitive to this need for interpretation. An article from their charter specifies:

> *7. We believe that it is important for the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology.*

5 Decision quality and prediction errors

5.1 Measuring the quality of a decision

In statistical learning, the accuracy of a decision depends on the quality of a prediction and therefore the quality of a model or algorithm. The latter depends on the *representativeness* or bias in the initial data, whether the model is adequate, and the quantity (variance) of residual noise. It is evaluated on an independent test sample or by cross validation (*Monte Carlo)* but remains indicative under its form as a *probabilistic error risk.*

Prediction methods are trained on learning data, so it is the quality of the learning data which is primarily decisive. Remember the old adage: garbage in, garbage out. The volume of learning data can be a useful factor of quality, but only if the data is effectively representative of the goal, and not biased. In the opposite case, terabytes do nothing. This was the case with Google Flu Trend (2008), which sought to follow in real-time and predict the course of an influenza epidemic, by drawing from a number of searches of certain associated keywords and with the knowledge of the location (IP address) of the user performing the search. It was the hype of the flu that was followed, and not the epidemic in itself. The data was taken, with better results, by a Boston team (Yanga et al, 2015), that considered a self-regression model that incorporated a hidden Markov chain that is corrected on the basis of Google search trends.

The literature on the subject of automatic learning is extremely prolific regarding ways to measure and estimate errors. It is firstly necessary to distinguish adjustment or learning errors, which define the proper use of the data, from errors due to forecasting or generalization. The type of measure or loss function is adapted to the type of the target variable or is provided for; this provision can be real, quantitative (regression problems) or discrete and qualitative (supervised classification). In a quantitative case, the loss function can be quadratic (mean squared error) or based on an absolute value, which is more robust than atypical values, but also more complex to optimize. In

a qualitative case, this can involve a simple error rate or an entropy measure, or more complex measures, as is particularly in the case in unequal classes. Learning error generally penalizes the loss function in order to control the complexity of the model with the value of the parameter to be optimized. The aim is to reach a better compromise between bias vs. variance, in order to avoid over-learning. Ultimately, once the algorithm has been trained and optimized on the learning sample, it is the estimate of prediction error on a sufficiently sized, independent test sample that provides an indication of the quality of an algorithmic decision.

5.2 The implications of algorithmic decision quality

It is worth noting that error prediction impacts the bias and discriminatory features of a decision (see section 3.5 on the recidivism score example) and influences the choice of a method or algorithm when seeking a better compromise between precision and interpretability. Although it is in many respects fundamental to be able to discuss the desirability of a decision (e.g., the consequences of a medical diagnosis), laws such as the GDPR make absolutely no mention of it. It would certainly seem very relevant that an algorithmic decision be accompanied by an evaluation of incurred error risk, just as the law requires pollsters to publish measures of uncertainty.

The main providers and vendors of Artificial Intelligence (Google, Facebook, IBM, Microsoft…) would be well advised to highlight and accentuate the most spectacular results of AI (image recognition, automatic translation, competition in the game Go…) with exceptional success rates, even better than that of human experts. But these successes are generally carried on the use of prototypes, or on applications where nothing is at stake. Unfortunately, the error rates associated with human behavioral prediction (recidivism scores of detainees, detection of insulting commentaries, fake news, risk behavior, etc.) are clearly, sadly or fortunately much more pessimistic.

The legislation does not codify an obligation of result, but any ethical practice or a manual of best design practices requires any creator— as is the case for doctors— to respect an obligation of means: that is, to make every effort in order to assure the citizen, the client, or the patient that the decision made will be the best possible given the state of knowledge. Moreover, the evaluation of error and the distribution of its causes effectively contribute to the consideration of the division of responsibilities. The obligation to publish or provide information on the quality of the algorithm used would be, as is the case of polls, an important factor in user accountability.

6 Conclusion

The gradual realization of the potential power of the automatic decision-making systems that use statistical learning techniques to exploit masses of data which are now ubiquitous in all sectors of activity (commercial, administrative, economic, industrial, medical, etc.) inspires as much hope as it does legitimate concern. We cannot exclusively rely on the responsibility of the actors behind these changes, nor on the dynamic on the front lines of machine learning research, to avoid misconduct or even the trivialization of the misuse of these techniques. Some of the risks notably include discrimination, the arbitrary nature of decisions in which neither the relevancy nor the responsible party are clearly identifiable, the possible outcomes of development purely guided by technical possibility, and on biases, which may even be involuntary, induced by the process of data collection that conditions the behavior of algorithms. The risks can also include points that were not addressed in this brief article: data confidentiality, risk of re-identification, and restrictions to competition.

The main difficulty results from the fact that seriously addressing these issues requires the possession of highly developed technical skills, in order to have a clear understanding of the functioning of algorithms, and keeping a critical eye on the discourse surrounding them. In addition, addressing these questions requires not only the aforementioned technical expertise, but legal, societal, sociological, and even political or philosophical expertise. The content of the debates on the subject and the analysis of even the most recent legal texts show that the challenge is considerable.

- Discriminatory practice towards a person is punishable by law, but it is up to the victim to provide the proof. Contrary to the United States, no legal text in France defines what could define or measure discrimination (DIA of the Villani report) towards a group.

- The obligation of transparency or explainability imposes at best that a human intervene to take responsibility for a decision, and it is only binding in the case of French administrative decisions which prohibit the use of self-learning algorithms without human validation or control. It is the same case for the sale of online advertisement.

- No legal text requires publishing data on, or informing users of the prediction quality or the rate of error associated with the use of a learning algorithm.

The ensuing technological disruption allows for all behavioral possibilities and practices, whether they are ethical or not. Issues of discrimination are the best regulated by law, but also the most complex to identify. The oft-cited example of predictive justice (recidivism score) shows that although the resulting decisions can be largely statistically biased and therefore collectively discriminatory with the use of certain criteria, it does not make it easy for a person to show that they have been wronged. Moreover, this example shows that data, learning algorithm bases, and their selective collection modes are both reflections of our societies and the main source of errors and biases.

In turn, this situation motivates fundamental research to define models and to build algorithms that will respond to these criticisms. Ongoing investigations consist of seeking a better compromise between different constraints, such as explainability and quality of prediction, bias reduction and data confidentiality. Verifying the interpretability and explainability of an algorithm or its underlying model and controlling its predictive qualities (for example on a test sample), in order to predict its potential collective or individual biases are complex tasks. At present, no single stakeholder can claim to be able to control algorithmic fairness. A plurality of opposing powers is therefore necessary. Which actors are likely to undertake these monitoring? Some are the public regulators: CNIL, DGCCRF (fraud reduction), Competition Authority, judges (French jurisdictions and the CJEU), but can they afford to do this? Other actors are from the private sector: collaborative platforms (Data transparency lab, TransAlgo INRIA, National Digital Council, Media (ProPublica in the USA), NGO Data (Bayes Impact), but they are only just beginning to get off the ground and are difficult to fund.

Do we need to go further than the principles established by the law for a Digital Republic and the Law of 1978, modified by law number 2018-493? At the moment, it is necessary to allow time for the application of these laws to measure their scope. It may seem premature, even though the effectiveness of monitoring devices remains uncertain. Additionally, how can we more precisely formulate the framework conditions in the use of different algorithmic methods, considering that the field of application (commercial, legal, medical, administrative, etc.) considerably changes the environment, and thus the terms of explanation?

In this still-vague context, it hardly seems relevant to once and for all turn to lawmakers, except to possibly claim, as is the case with polls, the obligation to display an error rate. Academic research on the subject has only been emerging over the last 2-3 years, and it is important to take a step back before imposing a specific rule to respect. Other norms are beginning to appear, that is, simple ethical rules and best practices (soft law), which could help to better understand the conditions of fairness and algorithmic transparency. Cathy O'Neil discusses the necessity of a Hippocratic Oath for data scientists, and the idea has been taken up by numerous groups and associations, including France's Data for Good, while a working party offered to sign a code of conduct on the ethics of data practices. The number of initiatives has been increasing, and it would be difficult to create an exhaustive list. It should be noted that European and American statisticians have had long-standing codes of best practice, but these texts cannot be adopted without deep reflection, for example on the notion of free and informed consent when it is collected online.

Finally, the use of AI in our daily lives requires the trust of users, which is difficult to grant to suppliers and sellers of technology when there are no controls in place. Another solution consists of offering different companies and primary stakeholders the issuance of an independent label testifying, following an audit, their respect of fair data use. This is a solution that has been proposed by the company ORCAA, created by Cathy O'Neil, and even the startup Maathics.