Academic director: Dr. Massimo AIROLDI

MASTER THESIS

Bias in algorithms & machine learning: origins, remedies & ethical challenges

Joseph GIGNOUX

March 2020 - June 2020

« *How far do we have to stretch the picture*
Before pixelating the human texture? »

Nujabes, Shing02 - Luv(sic) Part 3

ACKNOWLEDGEMENTS

I would like to thank my thesis director, Dr. Massimo Airoldi, for helping me throughout this project. With responsiveness, confidence and kindness, he helped me to face every problem I encountered and contributed to find the right solutions in order to obtain the best results for this research. Together, we constructed a relevant thesis and I had the opportunity to develop a fascinating and challenging subject that I cared about.

I would like to thank each expert interviewed: Dr. Bardolle, Dr. Boutalbi, Mr. Champion, Dr. De La Roche, Mrs. Lair, Dr. Masure and Mr. Pineau who are major contributors to this research. Each of them took the time to bring me, in a positive and benevolent manner, their knowledge on these topics. Through the sharing of their experience and constructive thoughts, I learned a lot and was able to develop my thinking on this project.

I would also like to thank my family and friends for their support all along this project, and especially my mother for her sound advice on writing.

ABSTRACT

Through the growing presence of machine learning algorithms in our environment, the world is gradually discovering the negative consequences they can cause. Justice, credit scoring, health, education, etc., no sector escapes the algorithms, and even the most critical cases for human lives are not spared. However, they can easily be mistaken: polemics around racist and sexist algorithms are multiplying. Algorithms and the data they contain are intrinsically linked to the past, even as they are used to influence and predict the future. Algorithms around us are seizing on society's discriminatory behaviours and they highlight, amplify and automate them. Beyond media polemics, bias is a complex subject that can manifest itself in very distinct forms, without the knowledge of its conceptors. However, even when they are identified, they are difficult to remove, and debiasing is a costly process with no guarantee of results. Moreover, ethics is a controversial subject, depending on individual value systems that are not universally accepted. Nevertheless, the fields of computer science and ethics are very different, and their actors do not collaborate spontaneously. This thesis seeks, both through the literature review and through interviews with different experts, to link the two fields, bringing together operational and theoretical knowledges, in order to provide synergies and a better understanding of each other. Our research seeks to understand how bias are formed, investigates the different remedies to fight against their emergence, and finally, discuss the ethical challenges raised by these issues.

#algorithm #bias #ethics #machinelearning #deeplearning #artificialintelligence #ai #fairness

TABLE OF CONTENT

INTRODUCTION	**p.1**
LITERATURE REVIEW	**p.4**
I. Origins & classification of bias	**p.4**
A. Dataset bias	p.5
B. Algorithm bias	p.14
C. Usage bias	p.18
II. Remedies to bias: leads & challenges	**p.19**
A. Big data: a wrong track ?	p.20
B. Corrective measures: opportunities & difficulties	p.22
C. Limitation of the corrective approach & other perspectives	p.27
III. Ethical challenges	**p.30**
A. Introduction to ethics	p.31
B. Methods of solving dilemmas	p.35
C. Ethical issues & avenues for reflection	p.39
THESIS CONTRIBUTION	**p.44**
I. Research design	**p.44**
A. Methodology	p.45
B. Experts	p.47
II. Analysis	**p.51**
A. Perception & condition of bias	p.51
B. Internal & external pressures	p.54
C. Dialectic of the recommandations	p.59
CONCLUSION	**p.65**
REFERENCES	**p.69**
APPENDIXES	**p.74**
I. Interview guide	**p.74**
II. Interview transcripts	**p.76**
#1. Dr. Bardolle	p.76
#2. Dr. Boutalbi	p.81
#3. Mr. Champion	p.84
#4. Dr. De La Roche	p.90
#5. Mrs. Lair	p.95
#6. Dr. Masure	p.101
#7. Mr. Pineau	p.109

INTRODUCTION

In 2016, a controversy broke out about an algorithm called COMPAS. This algorithm was at the disposal of American judges as a decision support tool to predict the probability of recidivism of a defendant. After a review by ProPublica, COMPAS was found to be stigmatizing towards black people. In the same year, Microsoft launched its Tay chatbot on Twitter, using machine learning technology: in just a few hours, the artificial intelligence was manipulated by Internet users and twitted anti-semitic content. In 2018, it was revealed that Amazon's recruitment algorithm, which automatically selects candidates' CVs, had a highly sexist behaviour. In 2019, Apple launched a credit offer whose amount are set by an algorithm. Within a few weeks, it was discovered that the algorithm had become sexist, offering women credit up to twenty times less than men of equal status. The biggest technology firms such as Google, Facebook, Twitter, IBM and YouTube have also been at the centre of media controversies because of failures in their algorithms that have had varying degrees of impact. The cases of biased and stigmatizing algorithms are very numerous and the examples cited have contributed to bring the debate on algorithmic bias and on artificial intelligence ethics to the forefront. Many researchers are now working to solve these problems, seeking to identify bias and address them at different levels. (Greene et al., 2019; Osoba & Welser, 2017; Bertail et al., 2019)

In order to address these complex issues, it is necessary to have a clear vision of what an algorithm is. The first notion of the algorithm appeared in the ninth century thanks to the work of the Persian mathematician Al-Khwârizmî. Originally, an algorithm is therefore not numerical and refers to a finite and unambiguous sequence of operations or instructions to solve a problem or obtain a result. Although algorithms were first invented to perform calculations by hand, they are now mainly used to be integrated into computer programs: it is this aspect of the algorithm that we will retain. Thus, we can compare an algorithm to a cooking recipe where we can distinguish three phases: first the ingredients, then the instructions and finally the finished food. These three phases correspond respectively to the input data, the calculation operations or instructions and the result provided. Algorithms vary greatly in complexity but always involve this same sequence of steps. (Osoba & Welser, 2017)

In recent decades, the development in calculation power of computers and the explosion of digital data flows have enabled a strong growth of algorithms in our environment. Indeed, big data has too much data to be processed by humans but it is easily exploitable by algorithms. Also, pioneering researchers in artificial intelligence, relying in particular on the work of Turing, have demonstrated that one of the major characteristics of intelligence is its ability to learn from experiences, and that these experiences can be digitized by digital data: « *Their efforts led to the formulation of learning algorithms for training computing systems to learn and/or create useful internal models of the world.* » (Osoba & Welser, 2017, p.5) Thus, it can be said that the meaning of the term « artificial intelligence » as we understand it today, that means in the sense of machine learning, ultimately designates an aggregate of more or less complex and interdependent algorithms seeking to simulate traits of human intelligence. Therefore, we will consider here that artificial intelligence is a field of study of algorithms.

We also need to look at the meaning of the word « bias » in order to better understand our study. Strictly speaking, the word « bias » means « inclination », more generally, it means a flawed or distorted reasoning. In psychology, bias generally refers to cognitive bias, that means a more or less unconscious arbitrary thought or partiality. In the field of algorithms and statistics, bias

refers to a lack of neutrality, which manifests itself in particular at the time of the result. (Friedman & Nissenbaum, 1996) That said, algorithms are nowadays used largely to make predictions and estimates based on numerical data, with the direct or indirect purpose of interacting with humans. The term bias can be defined as follows: « *Accordingly, we use the term bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate.* » (Friedman & Nissenbaum, 1996, p.332)

Today, algorithms and statistics have taken a predominant place in many fields such as health, finance, security, politics, marketing, etc. Thus, their growing presence also implies the growth of bias in our environment that affect us directly. Although the most widespread bias are related to empirical data, these bias can in fact occur during the three stages that make up the algorithm, that means in the dataset, in the instruction sequence or at the output stage. Moreover, the emergence of machine learning algorithms is making systems more complex through two axes. Firstly, machine learning algorithms, and especially deep learning algorithms, are extremely complex and their calculations are not always understandable by humans. Second, learning algorithms learn on data that can be biased by multiple factors. More generally, algorithms evolve in human environments, full of errors and constraints, condemning them to be imperfect and to produce errors. Therefore, it can be said that all algorithms are subject to the emergence of bias and that the more complex a system is, the more it is exposed: « *The possibility of algorithmic bias is particularly worrisome for autonomous or semi-autonomous systems, as these need not involve a human in the loop who can detect and compensate for bias in the algorithm or model. In fact, as systems become more complicated and their workings more inscrutable to users, it may become increasingly difficult to understand how autonomous systems arrive at their decisions. To the extent that bias is determined by the process of decision making and not solely by outcomes, this inscrutability may challenge the very notion of human monitoring for bias. And so while autonomous systems might be regarded as neutral or impartial, they could nonetheless employ biased algorithms that do significant harm that goes unnoticed and uncorrected, perhaps until it is too late.* » (Danks & London, 2017, p.4691)

Thus, the emulsion created by the notions of algorithms and bias leads us to a third subject: ethics. The harmful consequences that algorithmic bias can cause in our daily lives legitimately generate fear and mistrust. In this sense, introducing ethics into intelligent systems seems imperative to us.

The aim of machine ethics can vary and has been expressed from several points of view. In that sense, artificial moral agents has been defined in particular as: « *Which can follow ethical principles; which are capable of ethical decision making; which have « an ethical dimension »; or which are capable of doing things that would require morality in humans analogous to one common definition of artificial intelligence.* » (Cave et al., 2019) Then, three levels of ethical behaviour were distinguished for the purpose of ethical machines: « *First, « implicit ethical agents » are machines that are constrained to promote ethical behavior, or at least avoid unethical behavior. Second, « explicit ethical agents » are (in some sense) able to represent or reason about ethical categories or principles. Third, a machine counts as a « full ethical agent » if it is comparable in many or most relevant respects to human moral decision-makers.* » (Cave et al., 2019) However, the definition of ethics is plural. It is a complex, intangible notion, linked to the singular values of individuals and depending of the socio-cultural context. Nevertheless,

making machines ethical means first of all that they must not harm humans, therefore, researchers try to find consensus to arrive at this point.

The aim of this study is to identify the main causes of bias in algorithms, to study the means and current avenues of reflection to remedy it, and finally, to analyse the ethical challenges raised by this dynamic.

In order to give the best coverage of our subject, the literature review constitutes a state of the art of the different approaches, combining the fields of computer science and social science.

The thesis contribution is a discussion of seven interviews conducted with experts on the subjects of algorithmic bias and their ethical issues. Our results will be presented in the form of a synthesis and a critical analysis of their comments, in relation to our literature review.

How are algorithmic bias formed, how can we resolve them, and what are the ethical challenges that emerge from them?

LITERATURE REVIEW

I. Origins & classification of bias

The need to identify and classify bias in an attempt to deal with them in the best possible way is obvious. Thus, in the field of computer science, many researchers and data scientists propose thematic classifications of the different bias. Here, we have tried to draw up a synthesis of the different approaches.

However, it is impossible to propose an exhaustive list of bias, especially because their very nature is that they manifest themselves unwittingly, since the knowledge we have of them is experimental. The computer science community regularly discovers new bias that creep into algorithms, whether they are statistical, psychological, social or practical in nature, manifesting themselves at varying levels of granularity. It would therefore be dangerous to pretend creating an exhaustive list of bias, since claiming to know them all would mean reducing our vigilance regarding them, thus creating a situation favourable to their emergence.

We have tried here to adopt a chronological approach to the appearance of bias: first the bias present in the datasets, i.e. the inputs; then the bias linked to the algorithm itself, i.e. during the algorithmic process and the production of the outputs; and finally the bias linked to the use of the algorithm.

It is important to keep in mind that a bias can be related to several categories depending on its level of granularity, or even to several other bias depending on the approach we use. We must therefore recognize that the classification we are making is arbitrary and is mainly intended to propose an assimilable and synthetic classification.

A. DATASET BIAS

Dataset bias can be numerous because artificial intelligences require a large, smooth, representative and adapted dataset. Thus, many bias can slip into the training dataset of the algorithms. (Roselli et al., 2019) In our chronological approach, we can distinguish four moments of appearance of the bias: first the *anterior bias*, i.e. bias pre-established before any creation process; then the *individual bias*, i.e. bias linked to the human subjectivity of the algorithm conceptors (by « conceptors » we mean all the algorithm designers such as developers, engineers, programers and data scientists); then the *statistical bias*, i.e. bias linked to weaknesses in the mathematical approach; and finally the *emerging bias*, linked to the durability of the dataset in time.

Anterior bias

- Empiricism

Pre-existing bias are mainly due to social prejudices, morals, customs and social habits. A computer system developed in a certain social context is likely to reproduce its bias. These bias can be introduced consciously and voluntarily by the creators of the system, but also unconsciously, even with a willingness to avoid such errors. For example, *societal bias* can be described as prevalent and entrenched prior to the design of the system: they are inherited from organizations, culture or institutions. (Friedman & Nissenbaum, 1996) This bias gives rise to the *historical data bias*, which stems from the fact that most algorithms learn on the basis of empirical data, biased by the human society who transmitted its prejudices, its errors, etc. Thus, these errors are inevitably transmitted to the system if the necessary adjustments are not made. (Roselli et al., 2019) For example, an algorithm that edits personalized insurance contracts based on empirical data, where historically men were privileged over women, will be favourable to men and unfavourable to women. Here, the bias arises from a historical bias that is itself inherited from a societal bias.

Osoba and Welser (2017) address the algorithmic bias of artificial intelligence from a general angle, notably through the data bias. For them, artificial intelligence is condemned to reproduce human bias because it learns from empirical data, which themselves carry numerous social bias. They also claim that giving an algorithm a biased dataset, i.e. unadjusted, can even be useful to better perceive and become aware about the bias in our society that usually manifest themselves unwittingly. (Osoba & Welser, 2017)

The reproduction of gender bias by algorithms is an excellent example of the consequences of empirical data on our systems. Leavy (2018) was interested in the algorithmic bias that create sexist discrimination against women. Here, the bias come from empirical data, through English texts composing the machine learning datasets, which thus reproduce these same bias. She was able to distinguish several kind of bias.

- The *naming bias* comes from the fact that some similar entities are not named equitably. For example, historically, language privileges men and stigmatizes women: « women » are more often called « girls » than « men » are called « boys » when referring to an adult. Also, the status « Mrs. » or « Miss » depends on the marital status of the woman, whereas men are always expressed as « Mr. »: women are therefore often named according to their relationship to men. Some expressions such as « family man » have no common equivalent such as « family woman », conversely « single mum » or « working mother » have no common equivalent

among men. We will use the terms « he » or « him » to express a group composed of women and men, with the masculine overriding the feminine. Finally, in order to refer to a woman in certain professions, a prefix such as « female lawyer » or « woman judge » is customary, reflecting a social exception. (Leavy, 2018) It should be noted, for example, that most online translators translate « nurse » (neutral) into « infirmière » (female) from English into French.

- The *order bias* refers to the general tendency to list items in a list in a common and arbitrary order. This can be from the most powerful to the least powerful, as for example « doctor and nurse », men before women as in « men and women » or « boys and girls », and sometimes the reverse as for titles of nobility in « ladies and gentlemen ». This hierarchy of enumeration is not random and, according to the work of Leavy (2018), gender is its the biggest influence. (Leavy, 2018)

- The *description bias* comes from the fact that some similar entities are not described fairly. For example, men are more often described in relation to their behaviour while women are more often described in relation to their appearance or sexuality. In addition, adjectives used to describe women are fewer and more pejorative than those used to describe men. Also, studies on the word « girls » have shown that they are more often described in a negative context and are more objectified than « boys ». (Leavy, 2018)

- The *metaphor bias* comes from the fact that some similar entities are not portrayed fairly. For example, there is a general tendency to use more numerous and more pejorative metaphors for women than for men. (Leavy, 2018)

- The *presence bias* is due to a significant difference in the frequency of occurrence of similar entities in a large volume of data. For example, in business literature men are mentioned ten times more often than women. This bias may therefore result in machine learning systems systematically attributing the business world favourably to men and unfavourably to women. (Leavy, 2018)

Thus, the presence of these textual bias creates stereotypes embedded in the datasets leading inevitably to the reproduction of the latter in the outputs of machine learning systems.

In addition to stereotypes, we can say that a first problem related to data bias is that it can significantly reduce system performance. For example, a team of researchers conducted experiments on natural language processing (NLP) algorithms to study the consequences of minorities in dataset: « *We have provided empirical evidence for the hypothesis that forgetting exceptional instances, either by editing them away according to some exceptionality criterion in memory-based learning or by abstracting from them in decision-tree learning, is harmful to generalization accuracy in language learning.* » (Daelamans et al., 1999)

- Outside world

At a lower level of granularity, other bias can weaken upstream datasets when data comes from the outside world. (Roselli et al., 2019)

- The *bias of individual samples* occurs in particular when the dataset contains personal data that cannot be publicly accessed, modified or verified. This can lead, for example, to a situation where the information is incomplete, inaccurate or outdated. The *bias of inaccurate data* originate from the fact that the learning data comes from outside. This is often the case with the collection of personal and confidential data, which is usually not « clean » and not

accessible for reprocessing or verification. Such data may be incomplete, false, corrupted, mistyped, etc. The automatic collection of these data, and their difficult verification, leads to the absorption of this altered data. (Roselli et al., 2019)

Thus, we can see that bias are widespread around us and are already embedded in the data before us. It then seems obvious to have to reprocess these data before using them, however we ourselves are a source of bias for the algorithms.

Individual bias

Individual bias are those introduced directly by the developer into the system (Friedman & Nissenbaum, 1996) that may manifest themselves voluntarily or involuntarily, particularly in relation to psychological arbitration. A distinction is made between *psychological bias* on the one hand and *intentional bia*s on the other.

- Psychological weaknesses

Cognitive bias come from individuals and influence algorithms. Humans are subject to weaknesses in their decision making that can be directly reflected in algorithmic bias. Cognitive bias are the result of a distortion in the processing of information: it is a deviation of logical thinking from reality. These bias, which generally occur without our knowledge, are very numerous and can appear in a wide variety of situations. For example, the *confirmation bias* can lead a conceptor to orient his algorithm to favour his vision of the world; the *illusory correlation bias* can lead a conceptor to believe in a dependence of certain variables that are nevertheless independent. All of these elements participate in influencing and guiding algorithms in an insidious way, transforming *cognitive bias* into *algorithmic bias*. (Bertail et al., 2019) We distinguish several bias emanating from cognitive bias:

- The *goal representation bias* is the first level where a bias may appear when defining the objective of the algorithm and making underlying assumptions. Often algorithm designers set goals that are too vague, or there are many possible approaches to creating the algorithm, allowing arbitrary assumptions to appear. This first step is a gateway for bias of various kinds. (Roselli et al., 2019) For example, an algorithm editing customized insurance policies explicitly asking for the sex of the client in order to integrate this variable into the personalization when it is irrelevant.

- The *surrogate data bias* comes from the fact that the data making up the systems' learning datasets must be digital, in very large quantities and easily accessible. Thus, when a desired information does not meet these criteria, developers can interpret other data by substitution to estimate the desired information, leading to errors and shortcuts. For example, using the credit score for a recruitment algorithm as a sign of « reliability » rather than letters of recommendation, because it is less convenient to analyze on a large scale, can lead to obvious errors because the association is not proven to be relevant. (Roselli et al., 2019)

- *Goal by substitution bias* can occur when one wishes to achieve a new goal by making assumptions based on an extrapolation of historical data. For example, if a video streaming platform wants to estimate its revenues by launching in a new country based on the figures of its previous launches abroad, it is exposed to many errors: the context may be different depending on the culture, the quality of the internet infrastructure, the age of the population, the urban and rural distribution, etc. Also, innovations in the platform or additional

functionalities may help to attract users who were not previously attracted to it. Thus, historical data should be taken with caution and can be a direct source of bias. (Roselli et al., 2019)

- The *bias of non generalizable features* arises from the obvious practical difficulty for developers to build large and properly labeled datasets. Thus, developers can rely on already existing and properly tuned databases for learning, but these datasets may be inadequate for the purpose of production use. For example, a news article text classifier using an earlier religious text classifier dataset will tend to emphasize irrelevant words such as « post » in their classification because of the wide distribution of this word in the learning datasets. (Roselli et al., 2019)

For Osoba and Welser (2017): « *With limited human direction, an artificial agent is only as good as the data it learns from.* » (p.17) Algorithmic processes are inflected by the human hand, which directly or indirectly transmits the bias of which it is the author. However, bias can sometimes be introduced voluntarily by humans for a predefined purpose.

- Conscious irrational choices

Developers may intentionally make irrational decisions in building their algorithms for moral, economic, practical, legal, or ethical reasons. These bias are not necessarily negative: they may indeed generate statistically false results in order to meet legal requirements for example. In particular, we can distinguish:

- The *algorithmic focus bias* which consists in arbitrarily passing or using certain available data, for moral or legal reasons for example. For Danks and London (2017), this bias results on « *The use of morally irrelevant categories to make morally relevant judgments* » (p.4693). The reverse is also possible by omitting a statistically significant and unbiased variable because the result would run counter to a moral norm. This type of situation is found in particular at the legal level when certain data are confidential (Danks & London, 2017) or when certain contents are protected by copyright. (Bozdag, 2013) The problem with this bias is that it constrain the creator to make a forced choice: to choose between a statistically compliant but deceitful algorithm or an upright but statistically biased algorithm. (Danks & London, 2017)

- The bias that we call *bias of exogenous factor* refers to the set of bias deliberately introduced into an algorithm in order to meet requirements outside the system. These bias can compromise the reliability of an algorithm for price or cost reasons, in which case we speak of e*conomic bias*, or for other reasons that may be legal, for example. This is notably the case of Google, which favoured its own services in its search engine and was condemned by the European Commission for abuse of a dominant position. Similarly, Facebook had been accused of having favoured the ads of its partners to the detriment of other advertisers on its platform. More generally, search engine advertising is often a source of bias because the algorithm favours advertisers' results. The latter are often less relevant than the organic results because they are based on far fewer criteria. (Bertail et al., 2019; Bozdag, 2013)

More broadly, many intentional bias can be found in algorithms, especially in *online information providers* (OIP). (Pitoura et al., 2017) These platforms are often biased and some processing operations are even done by hand by Google or Facebook. These OIPs have their own « editorial lines » and values: they make choices consistent with their beliefs and define whether a content is interesting, unsuitable for « community rules », inappropriate, offensive, etc. (Roselli

et al., 2019) In addition, these platforms must comply with the specific requirements of the various governments that submit to them requests that are also biased. (Bozdag, 2013) Thus, it is important to emphasize that the algorithms governing the most widely used platforms in the world are not neutral, since the outputs and interfaces are partly the result of arbitrary decisions.

Human subjectivity, collective or individual, permeates datasets with many bias. However, the statistical method is not infallible either and can also carry paralogisms.

Statistical bias

The *statistical bias* can be linked to the datasets but also come directly from failures in the use of the statistical method. (Bertail et al., 2019) A distinction is made between bias linked to the choice of variables and bias linked to the estimators.

- Choice of variables

For the same question, we can consider that there will be as many different answers as there are possible configurations of variables, and this sometimes means an infinity. The choice and weight allocated to each one directly determines the result, even though complex neural networks estimate and value billions of different pieces of information. We have thus identified two main bias related to the choice of the statistical variables to be included:

- The *bias of feature selection* comes from the choice of the features to be introduced as inputs to an algorithm, and which therefore determine the outputs. Indeed, depending on the choice of variables, the results can be completely different from one configuration to another. Thus, the omission of certain variables, because they are difficult to quantify, unavailable or arbitrarily discarded, directly conditions the result. (Roselli et al., 2019) For example, Facebook's algorithm favours multimedia content over links and status updates, or *Likes* has a weight four times higher than comments: these choices are completely arbitrary decisions that directly determine the configuration of the Facebook users' timeline. Another example is Twitter, which chooses hashtags as a « trending topic » based on two hierarchical characteristics: first novelty, then popularity. This system is criticized by some researchers who denounce an algorithm that favours viral content and that would favour an audience more attentive to novelty rather than to the debate on persistent problems. (Bozdag, 2013)

- The *omitted variable bias* occurs when a data item is not available, « forgotten », or difficult to quantify. For example, it is difficult to numerically measure an employee's level of leadership proficiency, resulting in incomplete analysis leading to false results. Algorithms rely on multiple variables, however, an omitted variable, whether independent or correlated with other variables, can be critical to the quality of a result. We take the example of an algorithm that determines whether an employee is a « good employee », « average employee » or « bad employee » based solely on his results and not on « soft skills » such as leadership. In the case where the variable « leadership » is omitted and independent of the results, then an employee with low results will be considered a « bad employee », even if he is a key driver for his team. Worse, if the omitted variable « leadership » is dependent on another, for example negatively correlated to the variable « results », then the algorithm will systematically consider an employee with good leadership and essential to his team as a « bad employee ». (Bertail et al., 2019)

The parameterization of the variables therefore directly determines the quality of the result of the algorithm. A second challenge is added to this, according to Price and Ball (2014): « *We rarely have access to complete data* » (p.10), so data scientists are obliged to create representative samples of the population to be studied. (Price & Ball, 2014)

- Sampling errors

The question of the proportion of each class represented in a sample is central. Several methods corresponding to different situations have been defined by scientists, (Hoste & Daelemans, 2005) but the risks of error are numerous.

- The *irrelevant correlation bia*s occurs when inappropriate characteristics are associated with the input data, creating inappropriate associations leading to misinterpretations by the algorithm. For example, an algorithm recognizing the gender of a person from a photograph of his face with datasets of long-haired women and short-haired men will interpret hair length as a determinant of classification. However, even though women are more likely to have long hair than men and vice versa, it would be incorrect for this variable to have an impact on the result. (Roselli et al., 2019)

- The *selection bias*, which very often occurs when creating an estimator or convenience sampling, happens when the characteristics of the study population are not representative of the general population. For example, on social networks, we often speak of *silent majority* and *noisy minority* in reference to the fact that a small number of users generate the majority of interactions (comments, likes, sharing, etc.) and a majority have a passive activity, acting as spectator: we call it *activity bias*. (Baeza-Yates, 2018) Thus, by selecting a certain population, one can omit another with different characteristics and draw biased results considered as a general truth. (Bertail et al., 2019) Seely-Gant and Frehill (2015) focused on the occurrence of selection bias in big data activities. This bias is problematic in particular because it can lead to false correlations because the sample is not random enough, highlighting relationships existing only in the sample or, on the contrary, excluding correlations existing in the overall population. The cause of selection bias in the era of big data is explained in particular by the fact that some people are better able to generate data than others: presence on social networks, access to the Internet, ability to use information technology, ability to buy computer hardware, online interactions, etc., are all elements that can define a particular population. Thus, the data making up the big data are generally not representative of the general population: the result is a *self-selected sample* situation. If, for example, we want to estimate the vote in the second round of a presidential election based on tweets favourable and unfavourable to each candidate on Twitter, we will probably end up with a biased result. Indeed, the sample of people active on Twitter is probably not representative at the country level: it is probably a younger population, urban, etc. This study will therefore ignore a whole part of the national population. In addition, the typical characteristics of a Twitter user could be positively correlated with the probability of voting for a candidate, thus accentuating a correlation that is specific to the sample and not generalizable, even though towards the overall population these characteristics could be extremely minority. However, many big data analyses are carried out on open-source platforms or social networks because the data are present in large quantities and are easily accessible, which strongly contributes to the propagation of this type of bias. (Seely-Gant & Frehill, 2015)

A certain number of pitfalls are then presented to the data scientist if he does not show excellent rigour and good hindsight when using the statistical method. Moreover, even an algorithm that is unbiased at the time of its conception can become biased after being put in production.

Emerging bias

Emerging bias are bias that appear after the production phase of the algorithm, even while it is in use. Here, we study emerging bias that jeopardize data integrity and that are directly related to the life cycle of the information composing the dataset.

- Reinforcement learning

Machine learning algorithms are self-modifying over time by learning from their post-production experiences: the reinforcement learning is both a great strength and a great risk for these algorithms. According to Shadowen (2017): « *Machine learning accomplishes an aspect of artificial intelligence in which some data from an external context may be ingested, understood and integrated into the algorithmic function to make predictions.* » (p.4) Thus, these a posteriori modifications may also allow new bias to enter the system, as the algorithm absorbs, considers and makes its own data from the outside world.

- The *censorship bias*, or *truncation bias*, are bias close to the *selection bias* that occur specifically in machine learning. These major bias come from the fact that « *we do not observe the whole probability distribution generating the data* ». (Bertail et al., 2019) This is very often the case in reinforcement learning used by machine learning: intelligent systems today interact with their environment and create their own experience by absorbing new data. More precisely, the system having learned about data sends it to us, we interact with it, creating new data that it will be able to absorb again. However, if the algorithm has as new data at its disposal only the past interactions with previously sent data, then we can say the algorithm will comforts itself in its own results, creating a *feedback loop*. This is for example the case for suggestion algorithms: the algorithm returns suggestions from past data, but must also return data of the « exploration of possibilities » type in order to complete the statistical information. Unfortunately, this second type of data is often associated with a cost whereas the first is directly linked to an obvious marketing benefit: these bias then lead to extreme and reductive simplifications. For example, in the case of a recommendation algorithm for an online bookstore, if one initially reads a few detective novels, then the platform will only offer novels of the same genre, and will not offer historical or philosophical books. The moment when this simplification borns appears then as problematic. (Bertail et al., 2019) Another case of this type is often blamed on social networks: they tend to comfort individuals in their « bubble of opinion », or « filter bubble », linking them to content and like-minded Internet users, thus limiting dialogue and threatening critical thinking. It is therefore noticeable that the implicit personalization of content by web platforms directly participates in this censorship bias. (Bozdag, 2013) This situation can create social fractures that can be summed up by the phrase « *You're right and everyone else is wrong* » (Baeza-Yates, 2018, p.55) when individuals are confronted with different opinions when they come out of their « bubble ». This « filter bubble » situation is often criticized in the media as damaging to social dialogue: this is accused of creating tribalism, fuelling communitarianism and accentuating ideological polarisations. The example that appears most often concerns conspiracy theories believers, who share many of their theories and opinions on social networks. Platforms, through their algorithmic functioning, actively participate in locking individuals into their bubble of opinion

and thus in confirming their beliefs because they are exclusively linked to a community and content that goes in their direction. This information filtering system can lead, in the most extreme cases, to social isolation on the part of individuals because it creates a strong gap between the world in which they evolve virtually and the outside world. In this sense, as companies like Facebook « *determines, according to its own interests, what we see and learn on its social network* ». (O'neil, 2016) It means for some people that, in some way, they are participating in controlling what we think, since thinking is directly linked to the information we perceive. To go further, some also denounce the fact that algorithms participate in overexposing certain sensational information, such as hate crimes, thus creating echo chambers.

- The *manipulated data bias* occurs when the algorithm is deliberately manipulated to harm it. This manipulation can occur either internally, through data poisoning, or externally, through reinforcement learning. (Roselli et al., 2019) For example, *Google bombing* consists in the manipulation by Internet users of Google's search algorithm in order to influence the results of certain searches. It works by creating several domain names and web pages containing a link in a phrase, for example multiple pages containing the word « idiot », all pointing to the official website of a public person. Then, the algorithm will make an association between the word and the targeted page, and Internet users who do the search « idiot » will systematically find the website of the person in question. (Zeller Jr., 2006) Other cases of artificial intelligences using reinforcement learning that have been made public on the web have been attacked by Internet users in order to distort their results: for example, Microsoft's chatbot called Tay has tweeted racist and anti-semitic content after exchanging with malicious Internet users. (Osoba & Welser, 2017) Another simpler example is the *Likes* metric on Faceboook: as the *Likes* are a direct indicator for the platform of the popularity of a content, it is very easy to call on « click farms » to artificially inflate the popularity of a publication. (Bozdag, 2013) These bias can occur without hacking or exploiting a loophole, as it uses the normal operation of the algorithm, which makes some machine learning algorithms particularly vulnerable.

The algorithm's ability to learn from its experiences can therefore backfire and deeply damage it. While data can be injured, it can also become obsolete or incomplete after the production phase.

- Data obsolescence

In some cases, if the dataset of the algorithm is not updated regularly after the production phase, then some bias related to the durability of the data may appear.

- The *bias of outdated data* comes from the fact that the data are not systematically up to date and may be out of date. This can happen, for example, when a developer downloads an online database locally for faster access. If the developer does not systematically update the database, for example because the operation requires time-consuming technical manipulations, then the data may be out of date and generate obsolete results. (Roselli et al., 2019)

- The *bias of new societal knowledge* comes from new knowledge in society that has not yet been incorporated into the computer system. For example, a computer system that recommends nearby addresses has not incorporated a new restaurant into the list of

> referenced establishments, thus generating incomplete results that may lead to an imperfect competitive situation. (Friedman & Nissenbaum, 1996)

If many bias are linked to the dataset that supports the algorithm and are therefore exogenous to it, the algorithm itself can also contain flaws and generate bias, even with a healthy database.

B. ALGORITHM BIAS

Algorithmic bias are intrinsic: they are rooted in its very nature as an algorithm and more specifically in the succession of instructions that make it up. A distinction is made between original bias, bias linked to internal functioning and bias coming from the results emitted by the algorithm.

Original bias

The algorithm is not a perfect tool and is not adapted to all situations, it has weaknesses that come from its very condition as an algorithm. There are three original bias of the algorithm.

- The *formalization of human constructs bias* appears when one wishes to make complex or intangible human constructs accessible to computers, such as emotions, semantics, intuitions, etc. For Friedman and Nissenbaum (1996), this bias comes into play when « *We quantify the qualitative, discretize the continuous, or formalize the nonformal.* » (p.334) For example, an algorithm advising a judge to assess a required sentence on the assumption that the personalization of the sentence is not subject to human or unequivocal interpretation can lead to obvious bias. This is because the algorithm works through numerical data and mathematical methods. The latter are not always adapted to the complexity and extent of human emotions and creations, which can lead to shortcuts, misunderstandings and errors.

- The *endogeneity bias* is linked to an intrinsic characteristic of the algorithm in machine learning: it learns on the basis of data from the past. However, some situations do not allow for being predicted by the past, at least partially, but by the future, whose modelling is sometimes extremely complex. (Bertail et al., 2019) Moral hazard is a good example of errors created through endogeneity. For example, in the case of credit scoring, an individual with a history of risky behaviour in the past may very well adopt exemplary behaviour when starting a family. Thus, it can be complex, or even impossible, for a credit scoring algorithm to integrate this type of variables which are difficult to predict and quantify: the individual's credit scoring is then not at all representative of the future behaviour.

- The *Garbage In, Garbage Out bias* (GIGO) states that no matter how good and complex an algorithm is, if the input data is biased then the output results will be biased as well. This bias is directly related to dataset bias: if the input data contain cognitive or discriminatory bias, for example, the outputs will inevitably be biased without proper processing of the information. Thus, without a favourable adjustment of the algorithm or the dataset, the latter will inevitably reproduce, or even amplify, the input errors. (Bertail et al., 2019)

Beyond the weaknesses specific to algorithms in general, the development phase of a particular algorithm can create additional bias.

Statistical processing bias

In the instruction sequence that is entrusted to an algorithm, certain bias may be introduced by the developers at the statistical level, either voluntarily or involuntarily.

- The *decontextualization bias* occurs when an algorithm does not treat all entities equally despite it considers all relevant factors to accomplish its task. For example, an algorithm that automatically selects the CVs of temporary agency workers corresponding to a given

assignment in alphabetical order of last names, will result in favouring people whose names begin with the first letters of the alphabet. (Friedman & Nissenbaum, 1996)

- The *algorithmic processing bias* is when the algorithm itself is biased. This happens in particular when using statistical estimators. Here Danks and London (2017) explain that a bias is not necessarily negative and that an algorithm can be deliberately biased, in the statistical sense of the term, to counteract another bias. This is for example the case of an algorithm that learns, which can reduce the impact of extreme values in its calculations for future performance: here the algorithm is deliberately biased to be more efficient. Algorithms are also biased to be more « human » or « ethical »: this is for example the case of a search engine that removes hate sites from its results. In this example, the algorithm is not statistically neutral, but deliberately biased to meet ethical requirements. (Danks & London, 2017)

- A common algorithmic bias is that of permanently taking into account a criterion that is useful for some uses but useless for others. For example, most search engines routinely use geolocation to customize results, however this is not always useful. In the case where geolocation is a useless criterion, then it may favor certain results at the expense of others more relevant. Here the consequence is twofold: the user gets lower quality results and the owners of the websites that match best are disadvantaged. Furthermore, this can lead to a systematic disadvantage of websites located in sparsely populated areas. (Bozdag, 2013)

- Algorithms can also fall victim to their own functioning if someone knows about them. This is for example the case of the algorithm of referencing of the search engine Google whose operation is increasingly well known. Thus, a developer can strongly optimize his website in order to improve its organic referencing, i.e., to be developed strategically to match the expectations of Google's algorithm. (Bozdag, 2013) These strategies, called SEO for « Search Engine Optimization », has even given rise to two opposing practices: *White hat SEO* and *Black hat SEO*. While the first approach is recommended by Google and consists in facilitating the work of the web crawling algorithm, the second approach is condemned by the search engine. *Black hat SEO* combines a set of so-called aggressive methods consisting in over-optimizing a website and trapping Google's algorithm in order to inflate its SEO. This method is considered abusive and as « cheating » by Google because it can lead the algorithm to strongly favor one site over another that would be more relevant. (Segal, 2011) In addition, many media on social network use similar methods called SMO for « Search Media Optimization », to promote their posts and grow their audience. Some of these strategies are sometimes criticized as unethical. For example, on Facebook, if a post quantitatively generates a lot of interactions, i.e., reactions (of the *Like* type), comments and shares, then the social network increases its visibility. Thus, some media are accused of writing their articles exclusively on the basis of the « hottest » topics of the moment, and using very cleaving headlines. This will generate a lot of reactions from Internet users wanting to defend their opinion, and create a more or less aggressive debates in the comments, thus increasing the visibility of the publication on the social network. These methods are denounced because they are accused of surfing on the Internet users' emotions and accentuating social tensions, especially since often the title of the articles is catchy but the content is quite poor. (Eustache & Trochet, 2017)

- The *non-coverage bias* occurs when an algorithm chooses to exclude certain data from its calculations. This bias mainly concerns algorithms that process very large datasets, especially in the emerging contexts of big data. For example, this is the case of search engines that

develop this bias through two axes. First, only numerical data are taken into account by these algorithms: all non-numerical information is therefore excluded. Second, search engines like Google do not index all web pages. This second axis can be explained by the fact that some pages may be similar, considered useless for the user, have a bad reputation, be voluntarily de-indexed by the owner, or have a too complex technical architecture that the algorithm is not able to exploit. (Bozdag, 2013)

- The *time drift bias* is intrinsically linked to streaming machine learning algorithms. This term refers to systems that operate through the collection of sequential data, available in streaming form and continuously analyzed. This situation allows for the emergence of bias because: « *The analysis of data over time windows of too limited duration has its limits and leads to ignoring certain characteristics of the studied phenomenon such as long-term trends, seasonal effects or breaks.* » (Bertail et al., 2019)

We note here that if some statistical bias can be considered as flaws, others can be introduced voluntarily by developers in order to fight against another bias: here the bias becomes a tool against another bias.

Also, other bias may appear when the algorithm produces a result.

Output bias

Even if an algorithm is correctly set up, some bias may occur when it produces results. This is often related to the fact that a developer cannot always be aware of all the uses that will be made of his algorithm, nor of all the situations it will be confronted with.

- The *interpretation bias* is the misinterpretation of the outputs produced by a computer system. More precisely, it arises from a discrepancy between the information produced by the algorithms and the information needs expected by the user. (Danks & London, 2017) For example, a natural language understanding (NLU) algorithm can analyze a speech and give an opinion on the emotions expressed, displaying for example the tag « Joy » after analyzing a political speech. However, one may wonder about the ambiguous meaning of this word: is it related to the lexical field? To the themes addressed? To the tone of the voice? Is the speaker joyful or does he try to convey this joy? Thus, the lack of sensitivity of a computer system can create bias of interpretation of ethical, statistical, or semantic origin for example.

- The *novel case bias* occurs when a machine learning algorithm is confronted with a class which it has not been trained on and produces a false response. For example, a plant image recognition algorithm is confronted with an unknown variety, and gives the name of the plant it knows that comes closest in appearance rather than responding « I don't know ». These responses are thus discrete but major bias, that can propagate on a larger scale if they are not detected. (Roselli et al., 2019)

- The *data learning bias* comes from a bias in the learning of the algorithm, either because the learning dataset is biased or because the training of the system is biased. This can occur without the developers even realizing it. *Overfitting* is a good example of data biased training: the algorithm does not have the ability to « step back » and generalize its learning to new situations. For example, training a recruitment algorithm on a selection of exclusively French CVs will prove ineffective for recruiting foreign CVs. (Danks & London, 2017)

Some bias are referred as *technical bias*, i.e. they result from the constraints and limitations of the technical or technological tools used. (Friedman & Nissenbaum, 1996)

- The *bias of computer tools* comes from the limitations of, for example, the hardware, software or peripherals used. For example, in the case of a search engine interface with a maximum of ten results per page, the results on the first page will always be favoured even if more than ten results have exactly the same level of relevance for the associated search. (Friedman & Nissenbaum, 1996) Researchers have also conducted experiments and have noted that for 50% of the queries on a search engine, users only look at the first three results. (Bozdag, 2013) In addition, one study analysed the attractiveness of search engine results based on the number of clicks according to the design of each result. This specific bias called *attractiveness bias* was defined by answering the question: « *How does attractiveness or perceived relevance impact click behavior?* ». This study was therefore based on the fact that human attention is focused on specific elements of the design, such as a bold title or a bold summary, directly influencing our perception of the result, and de facto our behaviour. Thus, the study showed that beyond the correlation with the relevance of a result, our click behaviour is biased by the stylistic attractiveness of a result. After adjusting the relevance and position factors, the perception of a result significantly inflates its probability of being clicked. (Yue et al., 2010) This bias shows the impact of the user experience (UX) and user interface (UI) research, and how the designers can influence behaviours related to algorithms.

This last bias is also linked to the users of the algorithm, who can also be the source of many flaws.

C. USAGE BIAS

Here we find *emerging bias*, which are a category of bias that appear a posteriori from the creation of the computer system. Usage bias are often the product of the dynamics of interactions between users and interfaces, linked to changes in habits, culture, morals or social issues. (Friedman & Nissenbaum, 1996)

- The *mismatch between users and system design bias* stems from a significant mismatch between the users targeted by the conceptors and the population that actually uses the system. This mismatch may be due to differences in expertise, capacity or values and may be related to a misuse of the system. (Friedman & Nissenbaum, 1996) For example, a speech recognition algorithm that generates subtitles for people who want to watch a movie in its original version is used by people with hearing impairments: in this case the algorithm will not mention the sound effects of the movie (music, noises, etc.).

- The *data mismatch bias* arises from a significant discrepancy between the data used during learning and the data used after production. The most telling example is a poorly trained facial recognition algorithm on minorities: some of the most famous of those algorithms reached an accuracy difference of about 35% between a test on a light-skinned man and a dark-skinned woman. (Roselli et al., 2019) In this case, the consequences can be very serious. For example, a facial recognition algorithm used to identify wanted persons that is less well trained on a minority will be more likely to make a mistake on this population and thus confuse a face of an innocent with the one of a wanted subject.

- The *context transfer bias* occurs when an algorithm is used for a purpose that is slightly or significantly different from what it was designed for, which may lead to statistical, moral or legal bias. Thus, many bias categorized as algorithmic are ultimately bias originating from the users themselves in that they divert the use of the system from its predefined context. Danks and London (2017) point out that the line between a *context transfer bias* and a *data learning bias* may seem fine, but there is a difference between a biased dataset and a tool used out of context. Taking our example of the recruitment algorithm that automatically selects CVs: if the latter is sold explicitly to companies for French candidates, then its dysfunction on foreign CVs is not a *data learning bias* but a *context transfer bias*. Here we notice that Danks and London (2017) come close to the *bias of mismatch between users and system design* evoked by Friedman and Nissenbaum (1996).

Thus, these bias may be similar to data bias, however we have chosen a different approach in this classification because here the dataset is in accordance with the requirements of the developers. For them, they are not biased to the extent that they meet their expectations. Here these bias appear because the use made of the algorithm does not correspond to its intended use.

II. Remedies to bias: leads & challenges

The motivations for resolving algorithmic bias are multiple and are mainly based on ethical or efficiency issues. However, there are many bias, and relying on imperfect real-world data condemns algorithms to be imperfect as well. (Shadowen, 2017) Nevertheless, the search for remedies is not in vain, insofar as conceptors still have an interest in reducing bias even if they are aware that they cannot completely eliminate them.

In particular, big data is considered by some researchers as a good method, even a miracle cure, to fight against algorithmic bias. According to some conceptors, the use of a large volume of data in the machine learning process would make it possible to move towards greater statistical accuracy and to « drown » the bias because they would become derisory in the mass (Anderson, 2008). However, this theory is often contested (Thompson, 2019).

More generally, there is no univocal solution to the resolution of bias, but the best way seems to be a balance between technical and non-technical approaches. (Osoba & Welser, 2017) Several approaches are thus proposed to limit algorithmic bias. The most pragmatic ones recommend ways to directly deal with specific bias and are more adapted to palpable and identified bias rather than for less tangible ones, resulting in particular from complex human constructions.

We will first study the specific case of big data, then we will see the corrective methods for resolving algorithmic bias, and finally we will study the various reflections and difficulties beyond this approach.

A. BIG DATA: A WRONG TRACK ?

The challenges related to big data are numerous today and they are linked to the constraints of manipulating large datasets. More precisely, the question is: « *how to capture, transfer, store, clean, analyze, filter, search, share, and visualize such data* ». (Baeza-Yates, 2013, p.1) Even if the volume of data to be collected in order to be considered as big data is not clearly determined, in the digital age having large volumes of data is very easy. (Baeza-Yates, 2013) However, some researchers caution about the limitations of big data: « *Big Data offers great promise but also poses considerable risks. […] unfair discrimination is one of the most pressing, but at the same time an often underestimated issue in data mining*. » (Favaretto et al., 2019)

Big data & the « myth of large »

In this context, more and more researchers are therefore stating that the volume of data is not an end in itself, as it may be either useless or even generate difficulties and dysfunctions, (Dressel & Farid, 2018) but that we should rather seek to determine « *for a given problem, what is the right data and how much of it is needed* » (Baeza-Yates, 2013, p.1). Thus, researchers are trying to put an end to what they call « The myth of large », i.e. the belief that « *when data are big enough, bias are not significant* ». (Seely-Gant & Frehill, 2015, p.29) For them, a large volume of data does not systematically compensate bias thanks to its volume, contrary to common belief. In fact, like any data set and independently of its size, a big dataset can be biased by multiple factors. Contrariwise, the big data is very often the victim of the selection bias. (Seely-Gant & Frehill, 2015)

More specifically, researchers point out that big data is first and foremost composed of « found data », which means that it is based on the collection of available data and not on a rigorous and proactive statistical approach to collection, (McFarland & McFarland, 2015) thus echoing the problem of selection bias we mentioned, or the well-known problem of statistical sampling of convenience. Big data, which very easily exceeds the statistical requirements of a sample size, gives scientists a misleading feeling of confidence in their databases, even though the volume of the latter can make it even more difficult to detect certain bias. This problem challenges the common statistical belief that the more data there are, the more accurate and reliable they tend to be. Yet even as large volumes of data become easier to process, the same volume makes it very difficult to identify and fight bias. (McFarland & McFarland, 2015)

A context conducive to the emergence of statistical bias

The origin of the errors in the big data is that the large sample sizes result in extremely small variances and therefore statistically significant tests. This process results in a surplus of significant results that may seem very interesting to study, giving the researcher a false impression of high precision. The researcher is here a victim of cognitive bias in that he may tend to increase his confidence in his results rather than question his method of data collection. Furthermore, this cognitive bias is reinforced by the false impression of the statistical ideal generated by big data: the illusion of having a total population rather than a sample. (McFarland & McFarland, 2015) Yet, as we have seen, data collected in particular on the web can be highly biased, for example because of activity bias. This leads to a situation that can be described by the emergence of « *precisely inaccurate results* ». (McFarland & McFarland, 2015)

In addition, a major problem is that of data redundancy. Indeed, many high-volume databases contain the same data several times. For example, plagiarism content on the web is estimated at

25% and is linked to the fact that several websites contain the same texts or are translated into different languages. (Baeza-Yates, 2013)

Another problem is related to the disparity of data: this is explained by the fact that many algorithms follow a law of power. This means that extreme or uncommon data is discarded by the algorithms because it does not match the general trend of the data. However, there will always be cases where this information will be needed, even if it does not correspond to the normal distribution of the data. To illustrate, we can say that there will always be « *ordinary people with extraordinary tastes* » (Goel et al., 2010, p.1). For example, if individuals who are not considered as computer developers by a search engine are searching for information on a computer language by typing « python », they will only find results related to the animal homonym and not to the language, whereas the opposite will happen for a developer.

As a result, many researchers are now calling for « small data » to be preferred to big data in situations where it is possible, as it is considered easier to manage, to understand, more transparent, and sometimes just as effective. (Seely-Gant & Frehill, 2015; Dressel & Farid, 2018)

Also, to another extent, and increasingly aware of the risks associated with big data, some researchers are calling for the combination of quantitative and qualitative methods in research, a method that limits the risks of each approach and maximizes their respective benefits.

B. CORRECTIVE MEASURES: OPPORTUNITIES & DIFFICULTIES

Research related to « fair », liable and transparent machine learning is a growing field. Thus, many measures are recommended by researchers today (Osoba & Welser, 2017) and make it possible to provide developers with different tools and methods to limit the emergence of bias. In addition, more and more developers are developing quality control processes and imposing certain good practices on themselves in order to create systems with greater integrity.

Upstream resolution

- Statistical method

In the age of big data, data collection has changed profoundly from the collection methods used in the past. Thus, today we are more confronted with an observation of available data rather than a proactive approach to data collection. We can see on the one hand that big data allows quick and inexpensive access to large volumes of data, but on the other hand, the quality and source of this data is often obscure. Moreover, statistical bias are not necessarily compensated by the large volume of databases. (Bertail et al., 2019) Statistical methods can be used to compensate these bias by measuring or limiting them through the use of well-defined techniques:

- The most obvious method, especially in a big data context, seems to be the verification and adjustment of samples by population segmentation to fight selection bias. To address the imprecision of a biased sample, various segmentation techniques can be used to find a sample that is more representative of a population. For example, a population can be segmented according to the occupation of users or their place of work. It is also possible to use tandem segmentation, i.e., to integrate several characteristics. On the web, the most relevant segmentations are often related to the frequency, recency or value of a data item, especially to deal with activity bias. In addition, there are also techniques known as « community detection » techniques that make it possible to identify characteristic groups that are particularly active on part of a website. These techniques can be summarised as follows: « *If one segments found data by community detection and activity levels, then they have identified distinct populations at distinct observed activity levels, and analyses can be performed within each of the segments.* » (McFarland & McFarland, 2015, p.3) Thus, using segmentation methods, researchers can easily identify bias included in their sample in order to adjust them. (McFarland & McFarland, 2015)

- The *specialized learner method* consists of creating an additional learning dataset on its own to train a machine learning algorithm so that it can specifically learn the characteristics of a minority class. Specifically, instead of the algorithm being trained exclusively on the basic dataset, it is also trained on a separate dataset that contains data that deviates from the averages of the dataset as a whole. For a facial recognition algorithm, for example, this involves isolating certain minority demographic groups so that the algorithm can focus its training on the least represented populations. (Fuchs, 2018; Bertail et al., 2019) This works in three steps: the developer identifies the minority classes, then a specialized learner focuses on these particular classes, and finally the algorithm combines the knowledge of the specialized learner and the majority classes. Thus, the bias is corrected because the algorithm expands its range of possible values and its ability to focus on under-represented classes that would have been less well absorbed. (Howard et al., 2017)

- It is possible to « protect » the use of certain criteria in a dataset before starting the algorithmic process, i.e., to avoid that certain criteria are taken into account by the calculations of an algorithm. This notion of protected variable is even a legal obligation in some situations, in particular to avoid racial or sexist discrimination. In this sense, researchers have in particular developed tools to strengthen the statistical independence between the results and the protected variables. (Osoba & Welser, 2017) For example, it can be required that a criterion such as the sex of the individual should not be taken into account in the calculations. The limitation of this method is that it implies a statistical incompatibility between two notions: individual equity and group equity. In the case of a recruitment algorithm, individual equity recommends that each individual should be treated by the algorithm in a similar way, this is the concept of equal opportunity: for equal capacity, similar treatment is required. The group equity approach would recommend, for example, that women be given preference if they are disadvantaged. Here we note the incompatibility of the two approaches: it is statistically impossible to want to give equal opportunities to all individuals at the same time while applying a gender equity criterion. In our example, the problem can be summarized as follows: « *A problem with group fairness is that it does not take into account the individual merits of each group member and may lead in selecting the less qualified members of a group. On the other hand, individual fairness assumes a similarity metric of the individuals for the classification task at hand.* » (Pitoura et al., 2017) Another limitation is that a protected criteria can be correlated to another criteria which is not protected. For example, race can be correlated to the residential location which can lead to discriminating results even if the first criteria is protected.

- It is possible to correct algorithmic bias related to inferences by entering false data or by modifying certain data in the training dataset. This method consists of compensating for the misclassification rate of a group by introducing false data to compensate for the error spread between the different groups. For example, following the COMPAS case, researchers used this method: « *If subjects from some race were frequently misclassified as a repeat offender despite not reoffending, that minority then received an amount of falsified data in proportion to the rate of misclassification compared to how frequently other groups were misclassified in the same way* ». (Fuchs, 2018, p.10) Here the falsified data compensate for the bias contained in the empirical data and other hidden bias, in order to restore a fair estimate of the probability of recidivism, now focused on the relevant criteria. However, it is important to specify that this type of method should be used sparingly, as excessive falsification of data will inexorably lead to inaccurate representations of the populations studied. (Fuchs, 2018; Bertail et al., 2019)

- In the same logic, it is sometimes recommended to use the *hot-deck method* which consists in « completing the information » when some data are missing or unavailable: « *it is a question of reconstructing them by means of an appropriate statistical model, ideally adjusted using data from a controlled experiment design* ». (Bertail et al., 2019) For example, the missing data can be completed using the mean of a known characteristic present in a similar population. Even if this method is rigorously defined, it also has its limitations because the auxiliary model could also be biased or unsuitable for the first one, thus aggravating the bias present in the basic model. (Bertail et al., 2019)

- If the situation permits, sampling by separating the different classes may be relevant if one of them has a weak characteristic. For example, Twitter is used to provide a representative sample of the general US population and it is noted that the population on Twitter is much

more urban than rural. Yet, even when correcting the sample, i.e., adjusting the representation of rural users to match their proportion in the general population, they still constitute a much smaller proportion of the absolute number of tweets. By choosing to test the two classes, urban and rural, separately, we could find the origin of the bias: if the geolocation algorithm works with the same level of precision in both, then the bias comes from the population data that constitute the input, if there is a discrepancy, then the bias is structural and can be related, for example, to the quality of the network in the different areas. (Johnson et al., 2017)

There are at least as many control methods as there are statistical bias. However, it can be observed that very often, in order to fight against a statistical bias, it is necessary to introduce another one to counteract it: this means that the use of such methods must be done with caution, at the risk of leading to less efficient or even inaccurate algorithms. However, statistical methods that reduce the discriminatory behaviour of the algorithms also make it possible to considerably improve their accuracy. (Fuchs, 2018)

Researchers have highlighted the complexity of the fight against statistical bias through a study analysing the different discriminatory bias in four facial recognition algorithms using deep learning. Their study showed that four factors related to the data could create a racial bias: « *Sub-population distributions: the population statistics for demographic groups; algorithm: the quality of the algorithm's representations across demographic groups; representative images: the subgroup's representation of the population of interest; imaging conditions: the imaging conditions directly affect the difficulty of comparing images.* » (Cavazos et al., 2019) In addition, they also showed that the statistical choices in scenario modelling also directly influence the performance of the algorithm. (Cavazos et al., 2019) This example demonstrates that for the same situation, the statistical bias can be numerous and intertwined, making their resolution particularly difficult.

Researchers have also made various tools and stated the best practices available to conceptors to limit the occurrence of upstream bias.

- Frameworks

Some researchers have tried to design toolkits that can be used directly by developers to limit the introduction of bias in the algorithmic creation process.

This is the case of the IBM project: *AI Fairness 360* is a good example of this approach. This tool, which comes in the form of a software, uses data review methods, creation processes and analysis on the different stages of algorithm development in order to highlight the risks of bias. For example, this tool makes it possible to identify discrimination against minorities or individual bias through statistical estimates. (Bellamy et al., 2018)

However, these tools are often created to correct specific bias in specific contexts: although they are a good way to address some flaws, they are not able to eliminate all bias.

- Enlightened use of information

In public big data libraries, an increasingly common good practice is to add to the dataset a quality charter. This notice gives additional information about the data: reliability, sources, descriptions of certain categories, collection conditions, processing method, etc. This

information enables those who use it to check the good quality of the data and to better understand the dataset obtained in order to make better use of it. (Tam & Kimb, 2018)

Some of these means are effective, but each has its limits and are often situational or specific and not very generalizable.

Downstream resolution

- Auditing algorithms

One obvious way to fight algorithmic bias is to test them regularly in order to detect flaws and then correct them, especially in contexts with a high moral impact. This approach is iterative: each detected bias is corrected retrospectively. Database experiments and periodic model validity tests are essential to allow conceptors to find the balance between internal and external validity of the algorithm, but also between accuracy and fairness. (Shadowen, 2017)

However, bias can be difficult to erase or late detected: companies are therefore not always in favour of this type of operation internally because: « *any public statement about this will also inevitably be an acknowledgement that the problem persists* ». (Courtland, 2018, p.360) Indeed, it would be an admission of weakness on the part of companies that would force them to admit, for example, that their algorithms may have been or still are discriminatory. (Courtland, 2018)

To overcome this, some researchers are exploiting the possibility of auditing the algorithms externally, i.e., carrying out analyses by testing them as a user. For example, by submitting similar CVs to recruitment algorithms and changing a single characteristic, such as gender, in order to detect whether the algorithm is sexist. (Courtland, 2018) Even if this approach does not allow to know the core of the algorithm's functioning, it is the best way for an external person to get close to it, on the one hand because with the knowledge of certain inputs and outputs it is possible to get a big picture of the functioning, on the other hand because with a large volume of tests it is sometimes possible to arrive at relevant analyses.

- Transparency approach

Here we call « transparency » of an algorithm the fact of making it public so that it can be audited by a community of Internet users. Algorithm transparency is *« openness in the processes and methods used in the collection, processing, compilation and dissemination of the statistics* ». (Tam & Kimb, 2018, p.580) This is an important criterion increasingly used by democratic governments, which publish large online databases. This transparency generates trust with users and allows better control over the emergence of bias in the processes. (Tam & Kimb, 2018) Thus, « *If you can't be right, be honest* » (Courtland, 2018, p.360) can explain the desire to make algorithms transparent and open for two main reasons: on the one hand, to make possible the understanding of an algorithm, reducing its opacity and increasing trust in the service, and on the other hand, allowing to a large community to detect potential bias.

However, there are some problems with this kind of approach. First, knowing all of the parameters of a model do not necessarily allow an understanding of how it works. Second, some algorithms may contain personal or confidential data that cannot be revealed to the public, and it can be hard or costly to anonymize it. Finally, knowing how certain algorithms work would make it possible to exploit them in an opportunistic manner. (Courtland, 2018) For example, knowing how a CV screening algorithm works would allow candidates to know how to turn their application into a successful candidate by over-optimizing their candidacy.

Moreover, the problem with this transparency is that it is at the heart of divergent interests. For example, the NorthPointe company which markets the US justice algorithm COMPAS refused to disclose its algorithm out for intellectual property concerns, making its audit impossible. US citizens are therefore unable to verify for themselves a potential inequity in the algorithm, which reinforces their image of opacity and mistrust towards them. These two visions are strongly opposed because a public audit provides explanations, increased control, trust, satisfaction and better detection of bias in the algorithms. But, on the other hand, algorithms are complex intellectual creations that are protected by trade secrets and can manipulate personal data. (Bertail et al., 2019)

In conclusion, the transparency approach is not desirable for all algorithms as it could compromise its legitimacy or the confidentiality of certain data, which could even run counter to certain legislation such as the GDPR. (Courtland, 2018) One way to solve this problem would be the creation of trusted third parties auditing the algorithms and assuring the public of their compliance. (Bertail et al., 2019)

- Personalization & UX

Algorithms that allow users to customize their usage at the time of use help limit certain bias. In the case of a search engine, for example, a customization of the outputs will have several advantages: it increases the relevance and individualization of the displayed results and it can also reduce technical bias such as those related to interface display limits. In addition, it reduces the structural bias associated with the popularity criteria of a content. However, this approach also has its limitations: user interests may vary over time and the filter must then be updated at the risk of missing relevant information. In the case of OIPs, researchers have shown that sometimes users themselves do not even know precisely what interests them: « *user's declared and actual interests may differ* ». (Bozdag, 2013, p.221) Moreover, the mere wording of a topic can influence the user, varying whether it is considered interesting or not. (Bozdag, 2013)

Some of these approaches are interesting, but the major problem with pragmatic approaches is that they are more concerned with correcting the symptoms rather than addressing the source of the bias, whose reflections on this subject profoundly question our relationship to algorithmic technologies.

C. LIMITATIONS OF THE CORRECTIVE APPROACH & OTHER PERSPECTIVES

System complexity

Some bias are very difficult to identify in a set of data, and are referred to as *hidden bias*. In addition, it is sometimes difficult to understand the associations made by a machine learning algorithm. This is particularly the case for discriminatory bias present in large bodies of text: a human does not notice them when reading them because they are difficult to perceive, but algorithms can create insidious associations. Thus, it is sometimes difficult to identify the source of the bias in order to try to correct it, or even impossible when the volume of data is large and the associations are too complex. (Fuchs, 2018)

Also, we notice that a significant number of bias come from complex algorithms, which we are therefore unable to identify, because the operations performed by the system are too complicated to be understand by humans. Machine learning algorithms, and more specifically deep learning algorithms, are very often opaque and linked to the notion of « black box », i.e., even if we know the inputs and outputs of an algorithm, we do not know what is really going on inside it because their functioning is too complex to be understood by a human being. Thus, it is highly likely that some of these algorithms generate bias whose sources we cannot untangle.

In this sense, the question of returning to understandable and simpler algorithms is coming up more and more often in the computer science community: it is called « explainable AI ». Thus, some researchers are also recommending to prefer « small data » as big data, which would be more than sufficient in many situations. For example, the American software COMPAS integrates no less than 137 variables to estimate the probability of recidivism of a defendant. Dressel and Farid (2018) then decided to test the level of precision of a linear algorithm that would take very few variables to achieve the same objective. Thus, in their most basic algorithm, the researchers integrated only the two most relevant variables, which are the age of the accused and the number of previous convictions. With this model, the researchers achieved an overall success rate of 66.80% compared to 65.40% for COMPAS. Their algorithm had a lower accuracy on some intermediate categories, such as accuracy by race or the rate of false positive and false negative by race of only 1% to 3% less depending on the case. (Dressel & Farid, 2018) Thus, many complex systems could revert to simpler algorithms and more reasonable data volumes because the more human-sized the machine is, the more we can understand and control it.

However, even when auditing algorithms, the resolution of bias can sometimes remain a complex issue. For example, during the audit of four deep learning facial recognition algorithms, researchers highlighted the difficulty or even impossibility of resolving certain bias. After determining the factors that cause bias, namely biased data and inadequate modelling, they concluded as follows: « *Each of these factors individually, as well as interactions among factors, can impact bias. Data-driven factors underlie real differences in an algorithm's capacity to recognize faces of different races. Scenario-based factors are part of the measurement process and affect our estimates of algorithm bias. Given the complexity of these factors, and their potential to interact, it is not possible to perform a general assessment of bias for face recognition algorithms. Consequently, race bias must be measured for each particular scenario, algorithm, race, and dataset.* » (Cavazos et al., 2019)

In response to this complexity of systems, some researchers recommend to adjust the balance between human and algorithm in decision making.

Human/system balance

For algorithms where it is applicable, a good approach, at least in the short term, is to balance the responsibilities delegated to the algorithm with human intervention in the decision-making process. In some processes, the human eye is in the best position to identify the various bias that may appear. Focusing on human intervention at certain stages of the algorithmic process may limit the speed of some operations, but it is a price to pay for not being dependent of algorithmic bias. (Raub, 2018) For example, for a recruitment algorithm, it would be necessary to regularly check rejected CVs or to use statistics to identify possible correlations between the probability of success and irrelevant or discriminatory characteristics.

Therefore, it is important to use various methods of bias control, first during the different stages of algorithm design, but also at the same level, testing the benefits and limitations of each approach. Moreover, testing different algorithms with a similar goal can be very relevant in order to evaluate the behaviour of each one. It can be said that: « *Only rigorous experimental comparisons together with a qualitative analysis and explanation of their results can help determine the appropriate methods for particular problems* ». (Mooney, 1996) For example, researchers have conducted comparative experiments by testing seven *natural language understanding* (NLU) algorithms on the same body of text and observed different bias in each. (Mooney, 1996) It is therefore necessary to take a step back from the algorithms and more specifically from the methods: testing several of them and analyzing their results can help create a better system, by learning the positive and improving lessons from each and applying them in the same entity.

The notion of balance between human and system is also recommended in the interfaces to be designed for users. Researchers recommend better anticipation and individualization of the different contexts in which the algorithms will be used, and more specifically those related to content personalization. The method they recommend consists in particular in integrating elements such as the social context into the interface design process through upstream philosophical and technical studies. For example, for a search engine, this consists in customizing the search results according to the user's profile (age, location, interests, etc.): this technique can have its advantages, especially if the user is informed about the application of these default filters, called *implicit personalization*, and, even better, if the user has the possibility to modify or delete these filters, called *explicit personalization*.

In addition, the researchers also recommend a much greater « diversity of exposure » in the face of « diversity of content ». This means that even if an OIP has a wide variety of content at its disposal, a user will only see a small minority of it because of implicit or intentional customization. In their view, it is the duty of OIP to foster this diversity of exposure to information at the risk of encouraging the emergence of « filter bubbles ». (Bozdag, 2013)

Also, some researchers call for caution about our dependence on machine learning systems. In high-risk contexts and particularly in cases of use related to medicine, algorithmic bias can be critical and directly affect the health of individuals. Specialists therefore state that the automation of certain health diagnoses is desirable if it is measured, and call for extreme rigour concerning the objectives of using algorithms, the interpretation of results, the parameterization of systems and their transparency. These recommendations are made on the one hand in the sake of the

relevance and effectiveness of medical prescriptions, but also regarding the ethical stakes, which are particularly high in this sector of activity. (Gianfrancesco et al., 2018)

Getting to the source

Much of the algorithmic bias comes, as we have seen, from empirical data directly related to the prejudices and stereotypes of our society. Thus, apart from algorithms, reducing societal bias would be tantamount to directly reducing the bias arising from them.

To educate society, inform and sensitize people about harmful prejudices, implicit or explicit, based on race, nationality, region, social class, sexual orientation or gender, among others. Making individuals aware of this problem is a first step towards eliminating it. (Rudman, 2004) Thus, all the elements involved in breaking down the prejudices embedded in society contribute strongly to the improvement of algorithms and indirectly serve computer science.

Paradoxically, algorithmic bias help us in this sense, in that they place society face to face with its own past and present flaws, they highlight and extrapolate its defects. Thus, some researchers recommend adopting the *causal approach*, i.e., identifying the causes that allowed a bias to emerge, thus making it possible to justify the results and bring more transparency to decision support algorithms. (Osoba & Welser, 2017)

Therefore, the bias and the approaches of resolution raise many ethical debates.

III. Ethical challenges

The rising of big data and algorithms technologies in our society redesign our ways of living, for better or worse, legitimizing the question of ethics in many situations, as Favaretto and colleagues indicate, ethics is central because « *We should not forget that Big Data analytics – understood here as the plethora of advanced digital techniques (e.g. data mining, neural networks, deep learning, profiling, automatic decision making and scoring systems) designed to analyze large datasets with the aim of revealing patterns, trends and associations, related to human behavior – play an increasingly important role in our everyday life: the decision to accept or deny a loan, to grant or deny parole, or to accept or decline a job application are influenced by machines and algorithms rather than by individuals. Data analysis technologies are thus becoming more and more entwined with people's sensitive personal characteristics, their daily actions and their future opportunities.* » (Favaretto et al., 2019)

In that sense, the COMPAS case, which involves a decision-support algorithm for American judges, is the most controversial case concerning the ethics of algorithms, both in the research field and in the mainstream medias. As it was anticipated in the introduction, this algorithm was created by the NorthPointe company and made available to several American judges to estimate the probability of recidivism of a defendant through 137 criteria. The algorithm was audited by the ProPublica organization and the study revealed that it favoured white people and was racist towards black people. Overall, ProPublica showed that the recidivism rates of black and white people were about the same, 63% and 59% respectively, and NorthPointe had used protected values regarding race. (Shadowen, 2017) The algorithm's stigmatization is ultimately explained by the correlation between race and other factors that lead to racial disparities. This translated as follows: « *The mistakes made by COMPAS, however, affected black and white defendants differently: Black defendants who did not recidivate were incorrectly predicted to reoffend at a rate of 44.9%, nearly twice as high as their white counterparts at 23.5%; and white defendants who did recidivate were incorrectly predicted to not reoffend at a rate of 47.7%, nearly twice as high as their black counterparts at 28.0%. In other words, COMPAS scores appeared to favor white defendants over black defendants by underpredicting recidivism for white and overpredicting recidivism for black defendants.* » (Dressel & Farid, 2018, p.1) This polemic has thus strongly contributed to highlighting the debate on ethics in artificial intelligence and algorithms.

We will first look at the main notions that animate the debate on ethics, then we will study the different approaches to resolving ethical dilemmas, and finally we will look at alternative ways of reasoning about this issues.

A. INTRODUCTION TO ETHICS

Researchers have highlighted the legitimacy and the genesis of the emergence of ethical questions in technological issues, which Mittelstadt and colleagues define as following: « *In information societies, operations, decisions and choices previously left to humans are increasingly delegated to algorithms, which may advise, if not decide, about how data should be interpreted and what actions should be taken as a result. More and more often, algorithms mediate social processes, business transactions, governmental decisions, and how we perceive, understand, and interact among ourselves and with the environment. Gaps between the design and operation of algorithms and our understanding of their ethical implications can have severe consequences affecting individuals as well as groups and whole societies.* » (Mittelstadt et al., 2016)

On the importance of ethics in artificial intelligence

Recent advances in intelligent systems have attracted the interest of the general public and the media, particularly on ethical issues. The autonomous nature of these technologies raises many questions about what it means to make a decision for a system, the ethical, moral and legal consequences of such decisions, the responsibility of systems, the impact of the new context in decision-making, etc. Thus, it can be said that our ability to answer these questions will directly determine the level of trust that will be placed in artificial intelligence and, indeed, its place and impact in our society in the long run. (Dignum, 2018)

Most people fear artificial intelligence because they share a common vision of Artificial General Intelligence (AGI), i.e., self-conscious artificial intelligence with superhuman abilities and high amount of features: « *General AI is the Hollywood kind of AI. General AI is anything to do with sentient robots (who may or may not want to take over the world), consciousness inside computers, eternal life, or machines that « think » like humans.* » (Broussard, 2018) The current state of research in artificial intelligence is far from such a technology, and these fearful reactions can be described as excessive. However, the existential questions raised by this vision of artificial intelligence have led the scientific community to question itself further on the subject of ethical decision-making by autonomous systems. (Yu et al., 2018)

However, the pursuit of ethics is a complex but imperative process. Shadowen (2017) summarizes the challenge in the pursuit of ethics as follows: « *Any method used to solve for unfair predictions with legal and ethical consequences, is an example of machine ethics. A perfect solution to bias output in machine learning may never be possible because of our obligation to rely on imperfect data and the fundamentally prejudiced real-world we live in. Additionally, optimizing total societal good over accuracy in an isolated context is a complicated task. However, it is a moral obligation for computer scientists and those who purchase and employ the use of machine learning algorithms to pursue output distributions that most closely reflect an ideal state of fairness for all impacted individuals. A number of solutions have been either tested or conceptualized in the field of machine ethics to solve for bias in four main categories: technical, social, political, and philosophical.* » (p.16)

From bias issues to ethical issues

For some researchers, ethic is: « *a normative practical philosophical discipline of how one should act towards others* ». (Cointe et al., 2016, as cited in Yu et al., 2018) In that sense, some

researchers have been working on the classification of ethical risks arising from algorithmic bias. Six categories emerged from this research. (Mittelstadt et al., 2016)

- *Inconclusive evidence*: algorithms draw probable but neither inevitable nor certain conclusions. Thus, they are used when more reliable methods are unavailable or too expensive. Consequently, the results from learning algorithms will always contain uncertainty and probability of error, which can lead to unjustified or false results. (Mittelstadt et al., 2016)

- *Inscrutable evidence*: « *When data are used as (or processed to produce) evidence for a conclusion, it is reasonable to expect that the connection between the data and the conclusion should be accessible.* » This means that the complexity of many algorithms, and especially black boxes, makes the understanding of the reasoning unintelligible, thus making it impossible to examine or even criticize it. This phenomenon can lead to the opacity of algorithms, which is an obstacle to trust and error detection. (Mittelstadt et al., 2016)

- *Misguided evidence*: is an intrinsic weakness of the algorithm that we have highlighted with the GIGO bias, i.e., « *the output can never exceed the input* ». We can therefore deduce this principle: « *Conclusions can only be as reliable (but also as neutral) as the data they are based on. Evaluations of the neutrality of the process, and by connection whether the evidence produced is misguided, are of course observer-dependent.* » (Mittelstadt et al., 2016)

- *Unfair outcomes*: the designation of an algorithm as « ethical » can be based on many subjective criteria, depending on the judgement of the observer of the algorithm and regardless of the rigour of the conception: « *An action can be found discriminatory, for example, solely from its effect on a protected class of people, even if made on the basis of conclusive, scrutable and well-founded evidence.* » (Mittelstadt et al., 2016)

- *Transformative effects*: the consideration of algorithms as ethical stems from the fact that they do not seem to cause any obvious harm, however, algorithms themselves change our view of the world « *algorithms can affect how we conceptualise the world, and modify its social and political organisation* ». This is the case, for example, of recommendation algorithms that can influence the vision of users, especially since they are based on personal data. According to some researchers, this phenomenon can lead on the one hand to a decrease in the autonomy of subjects, and on the other hand, contribute to the loss of control of individuals over their data, thus threatening their privacy. (Mittelstadt et al., 2016)

- *Traceability*: it can be difficult to detect and identify the cause of harm caused by an algorithmic action. The dilution of the traceability of the causes of a damage threatens the notion of liability. (Mittelstadt et al., 2016)

Since those weaknesses directly lead to the rising of ethical issues, we must however define ethics in order to understand well the challenges revolving around this notion.

Definition of ethics & emergence of the dilemma

In philosophy, so-called normative ethics is the field of study of ethics that seeks to evaluate the moral quality of behaviours, actions and persons, according to the values of good and justice. There are three major approaches to normative ethics:

- *Consequentialist ethics*: the thought of giving more weight to the outcome of actions than to any other consideration. As its name suggests, it is an ethic oriented towards consequences, which must be as moral as possible: it is also called « utilitarian ethics ». (Yu et al., 2018)

- *Ethics of conduct:* a thought that states that every action should be judged by its compliance, or non-compliance, with its duties and associated obligations. It is also called « deontology ». (Yu et al., 2018)

- *Ethics of virtue*: the thought that an action is ethical if its actor carries it out in accordance with some of his moral values. (Yu et al., 2018)

For example, some researchers have succeeded in formalizing these ethical rules directly into autonomous vehicles: « *Deontology, a rule-based ethical framework, motivates the development of constraints on the system. Consequentialism, a cost-based ethical framework, motivates the construction of the objective function. The choice of weights is guided by the concepts of virtue ethics and role morality to determine behavior for different types of vehicles.* » (Thornton et al., 2017) And, according to them, certain benefits have been derived from the application of these rules: « *The normative ethical theories of deontology and consequentialism provide guiding principles for responsible programming of autonomous vehicles. […] Making these connections can enable engineers working at the deepest levels of programming automated vehicles to connect their design choices with broader issues of societal acceptance.* » (Thornton et al., 2017)

But, more generally, we can see from the definitions that ethics is an equivocal subject that has different levels. Beyond ethical principles, the question of values then arises and add an extra difficulty. The problem of values is that they are very subjective, implicit, and strongly linked to the socio-cultural context of individuals. Thus, for some researchers, systems should make these values explicit and fully integrate those of all stakeholders: « *AI reasoning should be able to take into account societal values, moral and ethical considerations; weigh the respective priorities of values held by different stakeholders in various multicultural contexts; explain its reasoning; and guarantee transparency* ». (Dignum, 2018, p.1)

Furthermore, there is an ethical dilemma when a decision has to be made by an algorithm, but all available choices necessarily lead to a violation of one of the ethical principles. (Yu et al., 2018) Some researchers even consider that: « *If you want to be fair in one way, you might necessarily be unfair in another* », (Courtland, 2018, p.358) meaning that every choices carry an ethical dilemma. Moreover, applying the concept of fairness to statistics is a real challenge: computer scientist Avid Narayanan held a conference that presented twenty-one definitions of fairness in statistics, and he did not claim to be exhaustive. For other researchers, statistical inequalities do not necessarily reveal ethical issues but sometimes simply the difficulty of studying one group rather than another. (Courtland, 2018)

Ethics is a major issue as autonomous systems interact with humans, and must therefore respect human rights and adopt acceptable ethical behaviour. Thus, various approaches have been used by researchers: psychological, social, legal, technical, etc., to address this issue. (Yu et al., 2018) In fact, the ambiguity is that there is an obvious and common recognition of ethical problems related to algorithms, but there is no concrete ideological universal alignment with their resolution: « *One of the most worrying but still under researched aspects of Big Data technologies is the risk of potential discrimination. Although « there is no universally accepted definition of discrimination », the term generally refers to acts, practices or policies that impose a relative disadvantage on persons because of their membership of a salient social or recognized*

vulnerable group based on gender, race, skin color, language, religion, political opinion, ethnic minority, etc. » (Favaretto et al., 2019) However, when we talk about harm, which goes beyond physically injuring or killing an individual, the definition of the word « discrimination » is vague and controversial, as behaviours may be considered stigmatizing by some but not by others.

However, beyond the differences in thinking, it can be said that the primary and universal objective of ethical research is to not harm human. (Varshney & Alemzadeh, 2017)

Various researchers have investigated ways to resolve ethical dilemmas in order to help designer algorithms to create moral artificial intelligences.

B. METHODS OF SOLVING DILEMMAS

Explainability & interpretability

The opacity of the algorithms is one of the main criticisms levelled at them. Algorithms often contain a *black box*, i.e., a part of the process between input and output that humans are not able to understand: we therefore know the input and output data, but we cannot explain the mechanisms that occur between the two stages. This black box is especially present in deep learning mechanisms because learning methods are complex. (Bertail et al., 2019) This lack of explicability poses obvious ethical problems, especially in contexts with a strong moral impact: « If the details are hidden, it's also harder to question the score or to protest against it. » (O'Neil, 2016)

In response to this, more and more research is now turning to *explainable AI*, which consists of returning to algorithms that are understandable to humans, removing the black box to make the algorithmic process intelligible. The objectives of these transparent artificial intelligences are multiple. (Bertail et al., 2019)

- First of all, it allows to verify and explain the results, even to non specialists of artificial intelligence. For example, in the health field, this would allow a doctor using a treatment recommendation algorithm to understand its suggestions, and to detect bias if any. (Bertail et al., 2019)

- It also allows new knowledge to emerge. A new result predicted by an algorithm can be useful to science, so we must be able to understand the origin of this discovery. (Bertail et al., 2019)

- Another reason is simply legal, firstly to support the right to challenge in cases of suspected defect or discrimination, but also for its compliance with the law. For example, the General Data Protection Regulation (GDPR) holds that: « *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.* » (Bertail et al., 2019)

- Also, there is an obvious question of confidence: the explicability of an algorithm is the only absolute guarantee of its correct operation. Without explicability, there will always be mistrust of algorithms because no other solution can totally guarantee a complete absence of failure. (Bertail et al., 2019)

The problem with the explanatory approach is that explicability is often negatively correlated to performance: the most efficient methods are, the most complex and therefore the least explicable. Consequently, simple algorithms are often less efficient than deep learning algorithms to achieve the same objective. It would therefore be necessary to distinguish situations in which humanity can afford an opaque algorithm and when it cannot, especially in light contexts, or when explicability is not crucial. (Bertail et al., 2019)

Therefore, in order to contribute to the resolution of ethical dilemmas, researchers have developed several useful tools in targeted application scenarios. We will present a few of them.

Tools for resolving ethical dilemmas

- General ethical dilemma analyzer: GenEth

A recent approach includes involving ethics experts and using crowdsourcing to resolve these dilemmas. Researchers have created a tool called GenEth that proposes to analyze ethical dilemmas by involving ethicists in the process of designing intelligent systems in order to codify ethical principles. This helps prevent artificial intelligences from going outside a previously defined ethical framework by using representation schemes. Specifically, after analysis of the recommendations of the ethicists, the actions of the systems are assigned positive or negative weightings according to their compliance with ethical requirements. In addition, a second variable is introduced according to the scenario in which the action takes place, with the aim of weighting the level of importance of the ethical duty related to the action in question. (Yu et al., 2018)

- Computational model of Moral Decision-Making: MoralDM

Researchers have created the MoralDM tool based on the premise that ethics involves not only utilitarian rules, but also depends on moral factors that are generally cultural. MoralDM therefore consists of including both approaches in the decision-making process. This means that some actions can be decided on the basis of their consequences, and others are automatically prohibited regardless of the consequences: these are the so-called *sacred values*. For example, a murder is always morally condemned regardless of the outcome. Thus, the creators of MoralDM are based on two principles: first, the consideration of explicitly established ethical rules such as sacred values, and second, analogical reasoning based on similar cases that have occurred in the past. In a given situation, if MoralDM does not identify sacred values, then it applies the basic rules of utilitarianism and directs its action towards the most useful result. On the other hand, if MoralDM detects a sacred value then it will favour a deontological approach and prefer inaction to action. (Yu et al., 2018)

- Software model of Belief-Desire-Intention: BDI

A framework has been created by researchers to provide an ethical framework for decision-making. In particular, it is based on theories of good, law and ethics and on the subjectivity of individuals. This system works like a funnel that proceeds by elimination to determine the right action to take: « *To judge the ethics of an agent's own actions, the awareness process generates the beliefs that describe the current situation facing the agent and the goals of the agent. Based on the beliefs and goals, the evaluation process generates the set of possible actions and desirable actions. The goodness process then computes the set of ethical actions based on the agent's beliefs, desires, actions, and moral value rules. Finally, the rightness process evaluates whether or not executing a possible action is right under the current situation and selects an action which satisfies the rightfulness requirement.* » (Yu et al., 2018)

- Online public platform: Moral Machine

Another approach is the Moral Machine project launched by the MIT to explore the ethical dilemmas related to autonomous vehicles in the event of an accident. Indeed, these technologies are part of a context where the ethical dilemma is very strong because the consequences are crucial and it is unprecedented: humans do not have to choose who to « kill » or protect in an accident, because their reaction time is too long for these decisions to be

considered voluntary. However, autonomous vehicles do have time to react: developers can therefore code the decisions to be made in critical situations into the machines beforehand. The objective is to decide what behaviour an autonomous vehicle should adopt according to each possible scenario, thus implying determining who the machine should protect or harm: the passengers or pedestrians for example. Moral Machine is presented in the form of a website allowing any Internet user to vote and to decide in each critical scenario the behaviour that the machine must adopt. The researchers then used these results and analysed them through eight criteria: « *Saving more lives; protecting passengers; upholding the law; avoiding intervention; gender preference; species preference; age preference, and social value preference*. » (Yu et al., 2018) This project allowed the researchers to identify several trends:

- The majority of Moral Machine participants choose the option that saves the most lives;
- Participants prefer scenarios that save pedestrians and kill passengers when it is not their vehicle, but they prefer the opposite when it is their vehicle;
- Reported preferences do not correspond to actual behaviour observed on the roads.

Thus, the results are contradictory and the ethical dilemma is extremely difficult to resolve. In response, some researchers recommend answering to these scenarios randomly and not coding them, while others recommend that autonomous vehicles should not be driven in the same places as humans in order to avoid this kind of scenario. (Yu et al., 2018)

Limitations of pragmatic methods

- Failure & shades about the social choice approach

The failure of the Moral Machine can be explained in particular by its operation based on the concept of social choice: artificial intelligence is here designed to act in accordance with the decisions of society. Researchers speak of « *coherent extrapolated volition* » and « *bottom-up ethics* », but this approach is very limited in that there is no global and unanimous ethical vision. Moreover, it is known that individuals as a whole do not necessarily wish to spontaneously give their opinion on ethical dilemmas, sometimes because they do not have one. This approach by social choice is nevertheless justified in so far as the creators of algorithms are no more legitimate than others in defining what is ethical and what is not, and, indeed, they are not legitimate in imposing their vision of ethics on the users of artificial intelligence. Moreover, the famous *wisdom of the crowd* theory states that the best results are obtained by soliciting the opinions of a large number of individuals: the more one collects the opinions of individuals, the more one tends towards accurate and neutral results. (Baum, 2017)

Thus, we note that there is no unanimous vision of ethics, nor does there seem to be a perfect approach to achieving a neutral and common ethic.

- Impossibility to predict all use cases

Methods of resolving ethical dilemmas help us to move towards more just and less biased artificial intelligences. However, these pragmatic approaches are very limited because, as we have seen, there is no unified and objective ethical vision. Moreover, these tools are adapted to specific uses and are not generalizable.

In this sense, since algorithm developers do not know the full range of situations in which their creations could be used, some researchers recommend defining a broader framework to ensure ethics in systems. Indeed, uses and contexts are not always predictable, which leads us to reflect on more flexible ethical rules. (Yu et al., 2018) For Shadowen (2017): « *Unfortunately, there is no one-size-fits-all ethical formula we can insert into a machine learning model and ensure unbiased results. If we are to combat bias using ethics, it will have to be on a case-by-case basis.* » (p.10)

C. ETHICAL ISSUES & AVENUES FOR REFLECTION

Exogenous difficulties in the ethical debate

Various exogenous factors influence the introduction of ethics in intelligent systems. In fact, the Montreal Declaration, which advocates an ethical conception of artificial intelligence, is articulated around seven themes that make up the moral context: « *universal concerns; objectively measured; expert oversight; values-driven determinism; design as locus of ethical scrutiny; better building; stakeholder-driven legitimacy; and machine translation* » (Greene et al., 2019, p.2122).

- Business hypocrisy

A first point raised by Floridi is the growing business around ethics in artificial intelligence. According to *Algorithm Watch 9*, more than seventy recommendations have been published between 2017 and 2019. This phenomenon generates confusion and inconsistencies in the application of ethics to algorithms. Thus, Floridi fears the emergence of what could be called *ethic washing*, like the famous *green washing* present in the marketing field. According to him, this trend can be explained by six types of manoeuvres and pressures exerted on companies in the AI sector. (Floridi, 2019)

- *Ethics shopping*: « *the malpractice of choosing, adapting, or revising (« mixing and matching ») ethical principles, guidelines, codes, frameworks, or other similar standards (especially but not only in the ethics of AI), from a variety of available offers, in order to retrofit some pre-existing behaviours (choices, processes, strategies, etc.), and hence justify them a posteriori, instead of implementing or improving new behaviours by benchmarking them against public, ethical standards.* » (Floridi, 2019)

- *Ethics bluewashing*: « *the malpractice of making unsubstantiated or misleading claims about, or implementing superficial measures in favour of, the ethical values and benefits of digital processes, products, services, or other solutions in order to appear more digitally ethical than one is.* » (Floridi, 2019)

- *Ethics lobbying*: « *the malpractice of exploiting digital ethics to delay, revise, replace, or avoid good and necessary legislation (or its enforcement) about the design, development, and deployment of digital processes, products, services, or other solutions.* » (Floridi, 2019)

- *Ethics dumping:* « *the malpractice of (a) exporting research activities about digital processes, products, services, or other solutions, in other contexts or places (e.g. by European organisations outside the EU) in ways that would be ethically unacceptable in the context or place of origin and (b) importing the outcomes of such unethical research activities.* » (Floridi, 2019)

- *Ethics shirking:* « *the malpractice of doing increasingly less « ethical work » (such as fulfilling duties, respecting rights, and honouring commitments) in a given context the lower the return of such ethical work in that context is mistakenly perceived to be.* » (Floridi, 2019)

As a result, we are witnessing a resurgence of malicious and illusory ethical practices, serving the business interests of companies and creating confusion rather than the real pursuit of establishing true ethical guidelines.

- Pressure of public opinion

The various controversies surrounding algorithms, such as COMPAS or Amazon's recruitment algorithm, have brought the question of ethics to the forefront. Thus, researchers on this subject know that they are subject to public scrutiny when carrying out their work: « *Building a moral background for ethical design is partly about shaping public perception, providing the concepts through which artificial intelligence and machine learning can be understood. One goal for these envisioning statements is thus to generate the moral consensus [...] [which] already existed within the scientific community on nuclear and biological weapons: acknowledgment of a specific set of threats, and a specific set of people, tools, and ideas ready to respond. Yet the problems remain, in this view, fundamentally technical, shielded from democratic intervention. The problems themselves are to be solved by experts in the technical features of AI/ML systems.* » (Greene et al., 2019, p.2129) Thus, some researchers believe that the pressure of public debate influences and harms scientific research, whereas the ethical debate must be led by the experts.

In addition, the complex issue of algorithmic liability puts further pressure on designer algorithms.

- Algorithmic liability

We can define liability as « *the principle that a person or organization legally responsible for the harm must provide an explanation or compensation for the harm suffered* ». (Bertail et al., 2019) Thus, this notion of liability is often linked to the notion of guilt. In particular, liability has raised a key ethical question in algorithms concerning the intentionality of discrimination.

Indeed, algorithms can discriminate even if there is no intention of the conceptor, and even if he has protected the necessary attributes, this happens for example when a protected attribute is correlated with an unprotected attribute. In this situation, it is more difficult to establish liability when it is unintentional. In this sense, it is also not easy to determine who has to make good the damage if a harm has been committed. Thus, some researchers recommend creating a new ethical and legal framework to consider algorithms as « electronic persons ». (Bertail et al., 2019)

In this sense, the law generally requires that protected attributes should not be used in order to avoid discriminatory bias, and their use is de facto considered as intentional discrimination. If the intentionality of discrimination in an algorithm is established, then the developer is liable to the extent that the discrimination is, in many countries, punishable by law. (Bertail et al., 2019)

More generally, according to some researchers, a responsible algorithm necessarily implies an underlying human responsibility for an implementation of artificial intelligence in our society in the long term. Thus, beyond the pre-determined ethical provisions that an artificial intelligence must respect, responsibility should be considered as a fundamental position in algorithm research. (Dignum, 2018)

The question of liability is all the more important when an ethical dilemma arises, i.e. the algorithm is in a situation where it is impossible not to violate an ethical rule, and therefore impossible not to cause harm. A good example of this situation is the study carried out by Karapapa and Borghi (2015) on Google suggestions and its *autocomplete* function. Indeed, Google has been the subject of numerous complaints concerning search results or suggestions that are considered defamatory, offensive or contrary to intellectual property rights. Google finds itself here in a situation of fundamental ethical conflict between: « *Freedoms of speech and*

access to information on the one hand, and personality rights of individuals, under a broader right of informational self-determination, on the other. » (Karapapa & Borghi, 2015, p.261) The example of the United States is interesting because the constitutionally protected freedom of expression occupies a special place there, so research suggestions are now considered as speeches and are protected by this amendment: « *Along this line, it has been suggested that the US jurisprudence supports the view that algorithm-based decisions, and, in general, outputs of an entirely automated process, are protectable under the First Amendment, as long as the algorithm or the automated process communicates a message to an audience.* » (Karapapa & Borghi, 2015, pp.266-267) If the law protects Google for publishing content, Google is no longer protected when the information is considered factually false: whether Google is an innocent content publisher or intermediary is then a essential question in determining its liability. To the extent that Google's algorithm uses an algorithm with objective factors and that it cannot have control over each suggestion, then the firm can be considered a mere intermediary. However, case law in the United Kingdom considers two independent factors in determining whether a search engine is an intermediary or a publisher, i.e., whether or not it is responsible for the content and therefore liable for defamation where applicable. Either the firm has had a passive approach in the publication of the content, or it « *has either actual or constructive knowledge of the content of the publication, namely, in case of defamation, that they knew or ought to have known by the exercise of reasonable care that the publication was likely to be defamatory.* » (Karapapa & Borghi, 2015, pp.271-272) If either of these two conditions is not met, then Google may be held liable for defamation. (Karapapa & Borghi, 2015)

This example shows us the importance of responsibility in situations with particularly heavy and complex ethical conflicts. Here, by considering the algorithm owner as responsible, then it diminishes freedom of expression, freedom of access to information and even network neutrality or internet governance. This phenomenon is accentuated by the legal process insofar as, for fear of being condemned, a search engine such as Google could use particularly restrictive measures on its algorithm and censor even more strongly than the censorship required by the laws. On the other hand, without content regulation, the personality rights of individuals, personal data and intellectual property would be at risk. Thus, the supposed neutrality of the algorithm here does not solve the problem. (Karapapa & Borghi, 2015)

Overtaking the opposition

- AI to cure AI

Some researchers have conceptualized the idea of a superintelligence, i.e., an artificial intelligence that would far surpass human intelligence, particularly in areas related to scientific creativity, general wisdom and social skills. Even if this concept is often contested, these researchers envisage the emergence of such intelligences within a few decades. A super-intelligence would thus be capable of creating other artificial intelligences superior to those existing today in multiple fields to the extent that it would be highly more efficient than humans. For moral thinking, this means: « *To the extent that ethics is a cognitive pursuit, a superintelligence could do it better than human thinkers. This means that questions about ethics, in so far as they have correct answers that can be arrived at by reasoning and weighting up of evidence, could be more accurately answered by a superintelligence than by humans.* » (Bostrom, 2003, p.3) In conclusion, given that humans are the primary source of bias in artificial intelligence, a superintelligence might then be better placed than we are to answer ethical

questions about artificial intelligence. This superintelligence then appears to be the best way to solve the ethical dilemmas in algorithms for the future.

- Redefining ambitions

Algorithms have a bad image with the general public. This can be explained in part by their reputation for being complex, opaque and difficult to understand. This image is in line with the reputation of artificial intelligence on the more or less dramatic aspect of AGI staged in popular culture. Moreover, algorithms are mostly evoked in mainstream media when there is a controversy about a discriminatory bias: COMPAS for example, has strongly contributed to spreading this serious and negative image of algorithms. Thus, we can conclude that there is a real aversion to algorithms on the part of the general public: they are often presented as the « executioners » which will decide on our insurance, our prison sentence, and even our death, referring to the dilemma of autonomous car. The most common recommendation to prevent humans from falling victim to algorithms is then the most obvious: humans must keep their hands on their systems. At first, this recommendation seems logical to us because the idea of depending on an uncompromising, opaque, complex and potentially biased algorithm for an important situation seems disturbing and dangerous for society.

However, not all algorithms are biased, and, ironically, we are committing a stereotypical bias when we distrust all algorithms because some of them are biased: we reproduce the same logical error in the algorithms that we blame them for us. Moreover, as we have seen, algorithms cause aversion, in part because they seem incomprehensible and opaque to the general public. However, human judgment is not necessarily transparent: it is simply more familiar to us. In this, we can admit that there is also a cognitive bias in the general aversion to algorithms.

Also, we can say that there is a mistake that is often made about algorithms: our expectations of them are too high. Researchers are looking for the highest possible accuracy and we have a real fear about margins of error, false positives and false negatives. So, we often expect our algorithms to be perfect, ethical, accurate and free of bias: we are very demanding with their results. However, to expect the COMPAS algorithm to achieve 100% accuracy is a fantasy, as it will never be possible to completely predict human behaviour that is related to free will.

Also, we add that humans are not perfect either and our judgment is not flawless. Furthermore, in most situations, algorithms would have to be highly biased to achieve a level of bias similar to a human. For example, researchers conducted a study to find out which algorithm was the most efficient to detect faulty software components on a production line. They did it because there are many techniques in this field, but none dominates and prediction remains problematic. To their surprise, when testing the different algorithm models, they found that the choice of algorithm on efficiency has very little impact, in fact the largest performance gap was only 1.3%. However, they found that the main performance factor is the group of researchers: the gap is 31% according to the team in charge. The study sums up its results as follows: « *It matters more who does the work than what is done.* » (Shepperd et al., 2014, p.1) In this case, we can see that the human bias is clearly more important than the machine bias. Moreover, the human is the direct cause of the bias in the predictions. Furthermore, it has been shown that the difference in sentence assigned by two judges for similar cases, or even the difference between two sentences assigned by the same judge on similar cases but spaced out in time, can lead to considerable differences. This is because human judgment contains a significant noise

component, i.e., random results that deviate from the mean: algorithms have almost no noise if they are statistically correct. (Naughton, 2019)

In this sense, it seems that the goal of the perfect algorithm is an ideal that is impossible to reach on the one hand, and that is not the right goal on the other hand. In this, we mean that it is vain to wish for a perfect algorithm, we must wish for an algorithm that does better than humans. This goal is achievable and already achieved in many areas. Algorithms finally make far fewer errors than humans in many fields and tend to improve, even if they are not perfect, they are still superior to humans in many situations. Moreover, it is usually easier to detect an algorithmic bias than a human bias and thus to correct it. We therefore argue that even if it is not infallible, an autonomous car will always be significantly safer than a conventional car, (Naughton, 2019) but because of its irrational aversion, it is not certain that the population will accept them any time soon.

We recognize, however, that the public is partly right, in that the learning process of algorithms is based exclusively on data and past experiences, though, certain fields such as philosophy, imagination, creativity and ethics are based on very human qualities that surpass math, at least until today. Thus, and despite the cognitive power of machines, some fields are simply not suitable for their use, making doubts legitimate in certain areas.

Thus, it is also necessary to define the situations for which the use of an algorithm is appropriate, in the light of the ethical and legal context, and considering the degree of uncertainty: « *In some applications such as cyber-physical systems and decision sciences, machine learning algorithms are used to support control and decision making in safety-critical settings with considerable costs and direct harmful impact on people's lives, such as injury or loss of life. In other applications, machine learning based predictions are only used in less critical settings for automated informational products. Applications with higher costs of unwanted outcomes tend to be also those with higher uncertainty and the ones with less severe outcomes are the ones with smaller uncertainty.* » (Varshney & Alemzadeh, 2017, p.17)

In conclusion, and regarding the various elements discussed, for the moment, the most relevant approach seems to be the situational, meaning a *case-by-case* approach. By using existing tools, by testing algorithms, by trying to reduce their opacity as much as possible, always taking into account the « gravity » of the context, (Shadowen, 2017) and keeping in mind that the good algorithm must above all do better than the human rather than do perfectly.

THESIS CONTRIBUTION

I. Research design

The objective of our contribution is to study the point of view of experts on the various issues raised in the literature review of this Master thesis: formation of algorithmic bias; methods of resolution of bias and their limits; emerging ethical challenges.

This section presents the chosen methodology and the various speakers involved in this research.

A. METHODOLOGY

The method chosen is that of the qualitative interview. This method makes it possible to collect the opinions of the interviewees in their own words.

Posture

The goal was to adopt a balanced presence in the interview. This means standing back to give interviewees the time and space to fully develop their convictions, while at the same time being present enough to guide the interview, i.e., when necessary, to ask questions, refocus the discussion if off-topic and request for clarification.

Also, it was necessary to be particularly vigilant about the posture adopted by trying, at best, to adopt an active listening posture. This means not making value judgements and mobilize skills of empathy and consideration to create a positive atmosphere.

Course of events

The interviews were conducted without time limits, in order to promote and encourage the development of ideas. Each interview was conducted via the Zoom application, either by video or audio call, depending on the preference of each interviewee.

Also, each interview began with a reminder of certain formalities essential to the smooth running of the interview:

- Request the interviewee's consent to record the interview;
- Recording;
- Reminder that the objective of the interview is exclusively the writing of the Master thesis;
- Ask if the interviewee wishes to remain anonymous;
- Brief presentation of the interviewer's background;
- Reminder of the subjects and issues of the thesis, as well as the main themes of the questions to be followed;
- Reminder that the interviewee should feel free to develop his or her point of view for each question;
- Reminder that the interviewee may not answer a question if he or she wishes to do so;
- Ask if the interviewee has a question or wishes to add something before the questions start.

No interviewee refused to be recorded, refused to answer a question, or wished to remain anonymous.

Questions

Each question was written with the aim of obtaining a long answer from the interviewee, avoiding being too specific. The questions are simple and clear so that they can be easily understood. Each question is asked one by one.

Each interview began with the objective of asking exactly the same seventeen questions. However, in order to adopt a flexible and coherent conversation, some questions could be deliberately omitted in some interviews. This happened either because the interviewee had already answered them in a previous answer, or because they were not relevant regarding the evolution of the interview.

The questions cut across the three main themes outlined in the literature review: formation of bias, bias resolution and ethical issues. The questions have been grouped into four main parts:

- *Introduction*: ask the interviewee to present his or her background.
- *Bias & corrective measures:* origins and sources of bias; approaches of resolution; request for an example of a situation where the interviewee was confronted with an algorithmic bias.
- *Ethics & application:* definition of ethics; issues of its application; request for an example of a situation where the interviewee was confronted with an ethical challenge related to algorithms.
- *Conclusion & openness:* balance between humans and algorithms in the decision process; research issues on those subjects; opinion on long-term developments; request if the interviewee wants to come back to a point raised, clarify an element, add a notion or ask a question.

(Question grid in appendix: *I. Interview guide*)

Transcript

Each interview was recorded and then transcribed in an effort to retain as accurately as possible what the interviewees had to say.

The transcriptions were made in an exhaustive manner, while turning from oral style to written style: this meant correcting syntax and language errors and reorganizing certain formulations. These slight modifications make it possible to transcribe more faithfully the substance of the ideas of the interviewees and to make the speech more intelligible, both for the reader and for the analysis.

The recording of the shortest interview lasts 27 minutes and 13 seconds and the longest lasts 57 minutes and 15 seconds. These differences can be explained simply by the respondents' propensity to develop their answers.

B. EXPERTS

Seven interviews were conducted. In order to obtain relevant answers, the interviewees were rigorously selected. Profiles were chosed on the basis of their professional experience and/or academic path.

Furthermore, it is pointed that the subject of this Master's thesis includes distinct areas of expertise that are not necessarily as well mastered by the all the experts. The subject of algorithmic bias in the strict sense requires technical skills in computer science, whereas the subjects of ethics are more often developed by experts with more theoretical knowledge. Moreover, in order to obtain a broad and deep understanding of the subject, some profiles correspond to academic backgrounds, bringing specialized and rigorous knowledge, while others correspond to corporate backgrounds, bringing empirical and operational knowledge. Thus, most of the interviewees concentrate several of these characteristics.

All the interviews were conducted in French with people with a perfect command of the language (native or bilingual). The experts interviewed are presented below.

Doctor Frédéric BARDOLLE

- Presentation

Beyond his thesis in machine learning and his role as head of the digital services incubator at the French Ministry of the Armed Forces, Doctor Bardolle has a strong sensitivity to subjects linking ethics and algorithms. Indeed, he co-founded and directed two renowned associations: Algotransparency and Data For Good. The first one consisted in a project allowing a better understanding by citizens of algorithms and their bias, in particular the YouTube recommendation algorithm. The second association consisted in the creation and federation of a community of 1.400 data scientists developing projects with social and solidarity impact for associations, institutions and startups. One of its projects has notably set up the « Serment d'Hippocrate des Data Scientists ».

- Current professional position

- Head of the Digital Services Incubator, Ministry of the Armed Forces.

- Notable academic title

- PhD, Computer Science, Strasbourg University.

(Interview transcript in appendix: *#1. Dr. Bardolle*)

Doctor Rafika BOUTALBI

- Presentation

Specialized in data science research, Doctor Boutalbi works on topics related to supervised and unsupervised machine learning. Her current work at Trinov focuses on the management and optimization of waste collection and sorting routes.

- Current professional position

- Data Science Researcher, Trinov.

- Notable academic title

- PhD, Computer Science, Paris Descartes University.

(Interview transcript in appendix: *#2. Dr. Boutalbi*)

Mister Renaud CHAMPION

- Presentation

Mister Champion has more than twenty years of experience in the artificial intelligence sector. During his career, he has been a board member, entrepreneur, investor and advisor to various companies, governments and international institutions such as the European Commission. As an active member of the IEEE (Institute of Electrical and Electronics Engineers), he is strongly interested in the ethical, legal and social economic issues related to AI and robotics. He is currently leading research on artificial intelligence at emlyon business school.

- Current professional positions

- Founder and Executive Director, AIM (Artificial Intelligence in Management) Institute of emlyon business school;

- AI Expert, IEEE.

- Notable academic titles

- MS, Engineering, Probability and Signal Processing, University of Michigan;

- MS, Electrical Engineering and Computer Science, CentraleSupelec.

(Interview transcript in appendix: *#3. Mr. Champion*)

Doctor Guillaume DE LA ROCHE

- Presentation

Trained engineer, Doctor De La Roche currently works in the division developing autonomous vehicles and driver assistance software at Renault Software Labs in Sofia Antipolis. His job consists in particular in testing and controlling the machine learning algorithms present in autonomous vehicles. In addition, he is also involved in training companies on the ethical issues related to algorithms and AI.

- Current professional positions

- Validation expert, Renault;
- Trainer in ethical AI, Éthique-IA.

- Notable academic titles

- PhD, Computer Science, INSA Lyon;
- Executive MBA, Business Administration and Management, emlyon business school.

(Interview transcript in appendix: *#4. Dr. De La Roche*)

Mrs. Caroline LAIR

- Presentation

After working in a startup developing a vocal assistant based on the principles of *privacy by design*, Mrs. Lair then founded her own company called The Good AI. The goal of this company is to centralize key resources in order to advise and support organizations in the development of responsible artificial intelligence. In addition, Mrs. Lair is the co-founder of Women in AI, an association seeking to promote the presence of women in the AI sector. Furthermore, she led the TEDx conference « *Intelligence Artificielle : Cheval de Troie pour une société plus juste ?* »

- Current professional positions

- Founder, The Good AI;
- Co-founder, Women in AI.

- Notable academic title

- MS, Management, emlyon business school.

(Interview transcript in appendix: *#5. Mrs. Lair*)

Doctor Anthony MASURE

- Presentation

A confirmed academic, Doctor Masure wrote his thesis on program design and has carried out numerous works on the relationship between design, digital and their social issues. He is also the co-founder of several journals dealing with these subjects such as « *Back Office* » and « *Réel-Virtuel: enjeux du numérique* ». Doctor Masure is also the author of the book *Design et humanités numériques* as well as numerous publications in scientific journals such as « *Résister aux boîtes noires. Design et intelligence artificielles* » or « *Le design de la transparence : une rhétorique au cœur des interfaces numérique* ».

- Current professional position

- Head of Irad (Institut de Recherche en Art & Design), HEAD (Haute École d'Art et Design) of Geneva.

- Notable academic titles

- PhD, Design, Paris 1 University;

- Aggregation, Design, ENS Cachan.

(Interview transcript in appendix: *#6. Dr. Masure*)

Mister Karl PINEAU

- Presentation

PhD student in Information and Communication Sciences at UTC, Mister Pineau wrote his Master thesis on cultural content recommendation algorithms. He is the co-founder and co-director of Designers Éthiques, an association promoting responsible and user-friendly design. This association is notably the organizer of the renowned annual conference Ethics By Design.

- Current professional position

- Research engineer, Decalog.

- Notable academic titles

- Doctoral student, Information and Communication Sciences, Compiègne Technology University;

- MS, Information Architecture, École Normale Supérieure (ENS) of Lyon.

(Interview transcript in appendix: *#7. Mr. Pineau*)

II. Analysis

A. PERCEPTION & CONDITION OF BIAS

The definition of a biased algorithm differs according to the approach taken. From a general point of view, algorithmic bias is defined as a gap between the output expected from the algorithm by its conceptors and the actual and manifested output. A bias is a deviation, an unexpected and unwanted result. We will analyse the different elements evoked by the interviewees.

Birth of bias

As artificial intelligences are aggregates of algorithms, they may contain different biased algorithms, which may themselves be biased at different levels. From a more technical point of view, bias can occur at different stages of the algorithmic process: data bias, conception bias and usage bias.

- Data

We note that among the interviewees, data bias is the most evoked and is considered as the most important. This is explained in particular because the data represent the level of the algorithm where the human weight is the highest. Thus, some even consider that there is no biased algorithm, but only biased data, because the data is what feeds the algorithm: « *A biased algorithm does not exist: the algorithm is biased by the input data that are biased, it is the information contained in the data that generates the bias.* » (Dr. Boutalbi) Without data, the algorithm is an empty box. In this sense, some experts think that algorithms are naturally biased because most of the datasets used are composed of biased data: they are empirical, historical or obtained from the real world, and therefore created by biased humans. In fact, these data simply represent a snapshot of a piece of society at a given point in time, a manifest human trend, and not a representative sample of what we want to observe. It can also be said that the data come from the past, which can, in some situations, add an additional gap between the desired output and the output obtained. In this regard, Dr. Bardolle points out that algorithms are very conservative systems: « *In essence, machine learning algorithms are conservative because they reuse data from the past. For Google Translate, the algorithm is trained on the same books in different languages and then, when there is enough data, it goes into production and is able to recognize the learned patterns. However, since the data comes from the past, especially because the quantities needed are large and therefore have to be collected back in time, we will observe undesirable behaviors. Here we reproduce the human bias of the past. If an algorithm translates « doctor » into the masculine gender, it is because it has seen it in many texts, and human societies were like that, but as societies change, algorithms do not have time to adapt.* »

Besides the origin of the data, bias may also arise from the fact that the data used to train the model do not cover all the cases that the algorithm will face when it is put into production. Here, the algorithm is then confused when it is confronted with a new case. This problem is described by Dr. De La Roche as one of the major sources of bias: « *In the case of the supervised machine learning, we have to try to cover all possible cases with the input data, but we can forget a situation, something may escape us, and that will create bias in our algorithm. […] We have to try to identify all possible problems beforehand in order to have input data that covers all of them.* »

- Conception

Concerning the bias of the algorithm itself, it can be caused by its conceptors. This means that since the conceptor of the algorithm is a human, it is imperfect: « *Data scientists are actually computer scientists who can omit or relativize the risks of bias. They are human beings who have their own perception.* » (Dr. Boutalbi) There is several origins of this bias.

On the one hand, this may be related to the fact that the conceptor has cognitive bias or carries his own vision of the world. Then, he will project it onto the system he is developing and will consequently make arbitrary choices according to his beliefs and culture. On this subject, some people therefore see a social cause. According to them, the majority of programmers belong to the same social group: they are mostly young, western men. As a result, these individuals share similar values that often do not represent society in its diversity, the users as a whole, or the context in which they are used. This phenomenon is particularly noticeable when the same service is used by a very large number of individuals around the world: there are inevitably cultural differences between certain users and the traits of the conceptors that are manifested in the algorithms.

On the other hand, there may be bias related to errors in the application of the statistical method. In the case of the learning machine, this bias is often related to the choice of the data training model: there are many models with different characteristics that are not appropriate for all situations. In addition, unsupervised machine learning models are more prone to bias than supervised models.

- Design

The third level of bias is related to the uses of the algorithm. This last point does not depend on the intrinsic characteristics of the algorithm, so it is often left out of consideration, but it remains a very important notion. Indeed, algorithms, in order to interact with humans, must manifest themselves in an interface. Design choices by designers and user-related usage bias then bring additional bias. The « dressing » of the algorithm is indispensable and produces real consequences: « *Design is important because algorithms are worthless unless they are integrated into services that cannot exist without interfaces, whether they be vocal, visual or other. Bias can be embedded in these interfaces, and designers play a role in this.* » (Mr. Masure).

Once the processes of bias formation have been identified, it is necessary to consider their consequences and the issues surrounding the resolution of bias.

On the usefulness of fighting bias

The consequences of bias in algorithms on the real world are twofold: from a statistical point of view, they can create errors of result or decrease the efficiency; from an ethical point of view, they can have a discriminatory impact on individuals. The definition of ethics varies greatly between individuals: it can be attached to notions of values or morals and it can be common or individual. Often, it is defined as an ultimate goal to be achieved in our actions, or as a moral frame of reference guiding our behaviour.

- Double-edged repercussions

Statistical bias does not necessarily have ethical implications, even though it can have adverse effects on the world around us. Dr. De La Roche gives us an interesting example of a biased

situation with no ethical impact: « *There are plenty of applications on smartphones to do barcode recognition and compare the price of products between different competing stores. If there is a mistake, it's not very serious, the user may pay a little bit more but there are no human implications or lives at stake.* »

Bias with an ethical impact concern in particular the sectors that are critical for human beings: justice, health, education, access to employment, etc. A biased algorithm can influence a decision, or threaten certain human freedoms. Above all, it can disseminate and automate biased and discriminatory human behaviour. In the case of the COMPAS algorithm, for example, the designated danger is not only the fact of influencing a decision based on the skin colour of a defendant, but it is that skin colour formally becomes the rule of the algorithm and that it materialises, automates and propagates racist behaviour. Mr. Champion sums up this problem: « *I don't see AI as creating new bias, but it does highlight the diversity of existing bias.* »

In addition, there are differences of opinion regarding the human capacity to combat bias.

- Uncertain victory

According to some interviewees, debiasing an algorithm is doing « unnatural » work, since the algorithm is naturally carrying bias. In their opinion, it seems impossible to completely remove a bias: it can be attenuated, but never completely removed, just as a human can never be completely neutral. Thus, one should not adopt a « technopositivist » nor a « solutionist » posture at the risk of having an excessive confidence in algorithms, and then creating a situation particularly favourable to the emergence of bias.

For others, the main difficulty in combating bias is that it happens without our knowledge, but the treatment is not so complex. The most difficult bias to deal with is the one we don't see, or notice too late, after it has gone into production: « *When the bias is identified, while taking into account the cultural context and frame of reference, dealing with it will be relatively simple. The challenge is really to identify them, to see them, not to let them pass. [...] It is not so much the bias that are created, but rather the bias that are not seen because a priori we are able to deal with them.* » (Mr. Champion)

However, both cases come to the same conclusion: the first step in resolving bias is vigilance and systematic questioning of the designed algorithm. It is necessary to continuously track down the bias, seek to correct the most obvious ones and identify the harmful consequences that the algorithm could produce.

B. INTERNAL & EXTERNAL PRESSURES

There are a numerous obstacles that complicate the process of debiasing algorithms. These barriers, either internal or external to the algorithm, can manifest themselves at different levels. Moreover, we will see through the YouTube recommendation algorithm, the ambiguities revolving around a biased algorithm. This example is notably based on the interview with Dr. Bardolle.

The YouTube case

From an ethical point of view, some consider as a bias the attribution of a « bad » goal to an algorithm. This means that the *success variable*, also called « metric » or KPI, attributed to an algorithm may be unadjusted or based on irrelevant variables. These two elements are complex because on the one hand they rest on subjective judgements, and, on the other hand, they may result from differences of interest between the conceptors and users of an algorithm.

- Meaning of success

The « success variable » of an algorithm corresponds to the goal set for it. For example, in the case of the YouTube algorithm, the success variable assigned to the recommendation algorithm is the maximization of the time spent by users on the platform, i.e. the « watch time », but this goal is questionable. Thus, videos related to conspiracy theories or promoting extreme ideas have particularly high *watch times*, i.e., a high average viewing time of a video by the audience, especially because they are more radical and sensational. Therefore, this kind of content is strongly favoured by the algorithm. If we consider that this content is problematic, then we notice several problems: the algorithm favours desinformation, misleading and divisive videos, so the success variable « watch time » is irrelevant in this case; the large-scale promotion of this type of video is unsuitable for users inasmuch as it produces individually and collectively harmful consequences; spending a maximum amount of time on YouTube, moreover in an environment with potentially pernicious content, is in the company's interest at the business level but diverges from that of the user's well-being; considering the preceding elements, we note that the choice of the « watch time » variable as a success variable is questionable, subjective and out of step with the interests of users.

In the long term, other consequences may emerge. For example, on YouTube, the *creators*, i.e., those who produce content, are in competition with each other because they share an audience with limited attention time. As a matter of fact, if the algorithm favours videos that polarize society, then a creator adopting rational economic behaviour has an interest in producing this type of content, even if he or she does not have the personal conviction to do so. The operation of the algorithm can have an incentive effect on creator behaviour and create a vicious circle. As a result, the massive diffusion of ostracizing and polarizing ideas can have a direct impact on society and accentuate ideological cleavages between individuals.

Nevertheless, correcting this bias would be complex: it would require a lot of time, financial and human resources from YouTube, with no guarantee of results. It would require a strong understanding of human and ethical issues, but on the one hand, they are difficult to synthesize and subjective, and on the other hand, they are very complicated to measure for an algorithm. Moreover, promoting « good time » on the platform rather than « long time » is not in the economic interest of YouTube. Indeed, this would require a deep questioning of the company's business model, which is essentially based on advertisers, i.e., on the visibility of ads, which

therefore depends directly on the audience's attention span. Dr. Bardolle speaks about this bias as the most difficult to resolve: « *It seems to me that these are the bias that go against the service's business model. Even if we manage to make them explicit, they are complicated to deal with because we will have to go through complex chains of decisions that may call into question the organization's model. We are not going to know how to deal with them without weakening the company. […] We can find approaches that work, but in the end, if we really want it to work, we question the business model of the company, which is based precisely on this brain time. […] If you don't change the model, then it can't evolve. So I would say that the most complicated bias to deal with are those that oppose the culture of commercial services.* »

- Fit a square into a circle

Dr. Bardolle raises another issues through a good analogy, is that of a cafeteria for young children where an algorithm would set the menus. Let's imagine that an algorithm tests new menus iteratively every day and its success variable is « children finish their plates »: if the plates are empty, then the menu is successful, if not, the algorithm must improve the menus. This goal has been defined in particular because the algorithm can only take into account numerical data: here it can only weigh the children's tray at the end of the meal to check if they are empty or not. In such a situation, it is very likely that after a few weeks, the algorithm will only propose very fatty or sweet foods: burgers, fries, ice cream, etc. Children generally love these dishes, so they will finish their plates more than if they have vegetables. This picture highlights, on the one hand, the problems related to the definition of an unsuitable target, and on the other hand, the limits of the algorithm in its ability to measure and quantify complex things. Reasonably, we can consider that children should not necessarily finish their plates, but rather eat a balanced diet in order to be healthy. However, this is more difficult to formalize for a numerical organism, it is easier for it to integrate quantitative rather than qualitative conditions.

Also, Dr. Bardolle's work includes the Algotransparency project, a program that measures the number of times a video is recommended by YouTube. This project demonstrated that the platform can be considered not only as a host, but also as having an editorial role. Indeed, it selects, according to its own criteria, the content that will be massively recommended to billions of users. In such a quantity of video with such an audience, the algorithmic recommendation behaviors are decisive, yet they appear to be arbitrary. To respond to this, Dr. Bardolle would suggest the creation of a new status for these platforms, which are today considered only as hosting providers and not subject to constraints of editorial responsibility as are the traditional media. In this sense, a new status could be given which also makes these platforms responsible, not for the content they host, but for the content they promote.

Dr. Bardolle recalled that the basic purpose of recommendation algorithms was very noble. On YouTube, it was going to allow any user to get recommended elements related to his passions, even if they are very much in the minority: the algorithm was going to be an asset to promote, cultivate, learn and share our interests. However, for certain specific content, such as political, conspiracy, violent videos, etc., then a deregulated recommendation algorithm represents risks. In the case of science, for example, it is a risky decision to subject to the same rules conspiratorial opinions and demonstrations that are the result of the scientific method and of the paradigm attached to it. Moreover, measures have already been implemented by YouTube to dissociate in its algorithm different kind of content. For example, this is the case of the creation of VEVO channels for music artists, which follow different rules than other videos, preventing music videos from being « flooded » by amateur content. (Wu et al., 2019)

Obstacles & uncertainties

We are going to see the different levels at which certain obstacles to the resolution of bias in algorithms manifest themselves.

- Statistic level: opacity

The presence of « black boxes » in machine learning is a disturbance in the resolution of bias that is regularly mentioned. Black boxes prevent developers from understanding how their own algorithms work, and thus from checking their reliability: « *Creating clusters can take a day, but trying to understand how clusters are formed can take three months. I think that's the big problem with algorithms, especially in machine learning or deep learning: algorithms do things that we don't really understand.* » (Mr. Pineau) Two issues emerge from the black boxes.

On the one hand, some recommend a return to « *explainable AI* » in order to understand the mechanisms of the algorithms and to identify bias where applicable. Coupled with this, the transparency and documentation of algorithms to stakeholders or to the general public, like open source software, has a role to play in cases where it is possible.

On the other hand, some people think that the *black box* phenomenon dilutes the notion of liability: the fact that algorithms are becoming « autonomous », learning « on their own » and having reasoning that we do not understand, is breaking the chain of liability. As a result, no one feels responsible when an algorithm produces harm, so conceptors lose interest in the issue of bias. By finding the right ways to make conceptors feel responsible, they will feel more concerned about the consequences of their algorithms and become involved safeguards. On this subject, Dr. Masure shades the liability of the conceptors in the bias: « *We must remain vigilant when talking about « racist algorithms » or « sexist algorithms » because we cannot attribute such an intention to a machine. However, the conceptors of these algorithms are not deliberately discriminating either, their bias come from social constructs, mostly unconscious, and no one is devoid of them.* » (Mr. Masure) This means that even if we call the conceptors « responsible » because they transmit their own bias to the algorithm, it is not a personal accusation because any human would transmit their bias. In a way, anyone constructing an algorithm is likely to bias it.

- Business level: incompatibility

Some denounce the business dynamics associated with the design of the algorithms around us: usable prototypes must be quickly created, like the MVP (Minimum Valuable Product), in order to make the investment profitable as quickly as possible. However, having an ethical approach takes time: it requires asking questions, setting up processes, etc. Thus, this constraint of speed appears to be a first incompatibility with the adoption of an ethical approach which is time-consuming.

Also, the search for ethics is rarely profitable for a company, and this notion is then set aside by the conceptors: « *Ethics is not always considered.* » (Ms Boutalbi) Thus, an obvious obstacle is simply that there is no economic interest in taking ethical measures.

Moreover, the application of ethics would require a profound change in the business model of companies because there is sometimes a strong gap between these two goals. For example, if we consider that captology practices are unethical, for a social network that is financed by advertising and therefore seeks to maximize the attention span of its users, then there is an

incompatibility: « *Are the goals that we set ourselves as individuals or as a society the same as those that we have assigned to our algorithms? The problem is that the answer is often no.* » (Dr. Bardolle) For some, having an ethical approach would even imply abandoning the quest for profit: « *It even tends to say that to be ethical, companies would have to be outside the economic market.* » (Mr. Pineau)

Another difficulty related to the business model is mentioned by Mrs. Lair: « *It would be dangerous to ask an American company to provide France with a predictive justice algorithm, because our ethical codes related to justice are different, it would have to be produced locally. However, it is contrary to business dynamics because a company does not want to limit itself to its domestic market, a French or American company generally has the ambition to export its services abroad.* » Moreover, there is the question of the concentration of the number of users, which leads us to ask « *Who decides for how many?* » (Dr. Masure) because companies generally seek to maximize their number of user. These elements represent a high risk, and, according to Dr. Masure: « *A service with two or three billion users is too concentrated, which creates a global problem. One can also think about how many people are affected by the same values a service carries.* »

Also, these oppositions that we have mentioned directly raise the question of the place of subjectivity in algorithms.

- Ethic level: subjectivity

Subjectivity is a great difficulty in the application of ethics. All respondents agree that ethics is an arbitrary notion, between individuals on the one hand, and culturally on the other.

Also, the different definitions of normative ethics seem to be found in the vision of the interviewees: all of them focus on the consequences and impacts of algorithms, i.e., consequentialist ethics, but some of them also evoke questions of intentionality, means and efforts, i.e., deontological ethics. However, the only element present in all visions of ethics seems to be that of « *not harming someone* » (Dr. Boutalbi), which is in line with consequentialist ethics. Thus, and as Mr. Pineau points out, the ethics evoked in everyday language is clearer than the philosophical notion: it is consequentialist ethics. According to Mr. Pineau: « *When we talk about ethics in everyday language, I think it is just « doing good ». When we talk about a business that is ethical, it is a business that does good, or that gives the feeling of not being aggressive from a marketing or commercial point of view.* » This definition seems to be the most relevant here because it is the most obvious way to approach the issue of algorithms. Its factual approach significantly reduces subjectivity by focusing more on the results produced by an algorithm rather than on metaphysical and intangible issues.

Moreover, it would seem that, accross cultures, certain actions in defined situations are universally condemned or approved by the majority of individuals. In this sense, it can be recognized that there are, in the ethics applied to algorithms, certain situations that are free of cultural differences. Consequently, there are situations in which it is possible to make decisions that are accepted by the majority of individuals, and others where it is not possible. It is a question of identifying these « *common denominators* » (Mr. Champion) across cultures in order to design services that are suitable for the greatest number of people. However, we have to keep in mind that this idea, which is the core of cultural anthropological research, is very controversial: many researchers believe that there is very little in common between cultures, or

even that ethics is purely a social construction, making it impossible to harmonize and build a common set of ethical rules because of their disparities.

In that sense, the more the algorithm is available to a small and homogeneous group of people, the less cultural dissonance it will cause. Even if this appears, as we have seen, to be contrary to the business dynamics of companies, it would be preferable to have « *algorithms by cultural era, instead of having an imperialist ethic imposing itself on everyone.* » (Mrs. Lair) Dr. Masure brings to this problem a line of thought. According to him, one solution would be to: « *Document the value system of the cultural context in which the algorithm is embedded. A kind of grid, or explicability beyond the technical field, to create a matrix of values and to be able to situate the context in which an algorithm is disseminated. This would allow conceptors to better situate their actions and to better predict the consequences.* »

According to Dr. Masure, there is no such thing as a neutral algorithm, just as there is no such thing as a neutral human. The question therefore must changes from « *How to be objective?* » to « *What values are we going to put into it? Who will decide?* » In this sense, Dr. Masure also suggests drawing inspiration from « design fiction » or « speculative design », which consists of « *developing scenarios by extrapolating the consequences* » in order to determine the concrete impacts of potential decisions.

The question of ethical decision-makers then becomes a central issue. Several are mentioned by the interviewees, who may intervene at different levels of the decision-making chain.

C. DIALECTIC OF THE RECOMMANDATIONS

Through four axes of reflection, we will study the different solutions to the resolution of bias and the application of algorithm ethics. The objective here is to highlight the opportunities and limits of the different tracks mentioned and to understand the ecosystem in which algorithms evolve.

Involve & communicate

A first element of reflection is primarily the human factor: it is about diversifying the stakeholders of the artificial intelligence sector and choosing who should decide on algorithmic ethics.

- Decision-makers

Some recommend an external advisory committee to reflect on the various ethical issues related to algorithms: « *An independent, multidisciplinary committee, representing civil society in all its diversity. Researchers, academics, companies, engineers, individuals from civil society.* » (Mr. Champion) This committee could act either as a consultant, a legal authority or a certifying body. We are already seeing the emergence of this type of organization in France and abroad, as stated by several interviewees, with recent institutions such as the CNIL[1], IEEE[2], ARCEP[3], CERNA[4], AI Alliance[5], etc. These are independent structures that may have legal authority, and therefore the power to impose sanctions, or simply to encourage good practices and provide information.

In this sense, Mr. Pineau suggests drawing inspiration from what is being done in the field of bioethics: « *We also seek the opinion of people working in other fields who can enlighten the decision to be taken, such as philosophers, historians, sociologists, even theologians or people from civil society.* » For him, the opinion of the population is legitimate and necessary because it « *will have such an impact on the population that it justifies giving them a say.* »

Dr. De La Roche suggested integrating all the actors involved in the use of a given technology: « *When an application uses AI, we must try to identify all the actors involved. In the case of an autonomous car, there are many players: the constructor, the parts manufacturers, the people on the road, the seller of the car, the insurance companies, the driver, the State, etc.* » Dr. Masure also supports this argument by suggesting to draw inspiration from what is done in « codesign », which is a collaborative design method involving both the designers and the users of a service. For example, one method is to survey potential future users of a product. This allows to identify new use cases that would have been omitted by the conceptors and thus to complete the learning data of the algorithm. It also makes it possible to collect people's opinions on the behaviours that the algorithm should adopt in different situations, so that there is no one-sided position of the conceptors on these issues.

---

[1] *CNIL*, [Online, https://www.cnil.fr/professionnel (consulted on June 15, 2020)

[2] *IEEE*, [Online], https://www.ieee.org/ (consulted on June 15, 2020)

[3] *ARCEP*, [Online], https://www.arcep.fr/ (consulted on June 15, 2020)

[4] Allistene, *CERNA*, [Online], http://cerna-ethics-allistene.org/ (consulted on June 15, 2020)

[5] European Comission, *The European AI Alliance*, [Online], https://ec.europa.eu/digital-single-market/en/european-ai-alliance (consulted on June 15, 2020)

However, Dr. De La Roche also shades the capacity of such initiatives, as they deal with complex, diverse and recent topics: « *We can't expect a miracle solution that will be unanimously agreed upon and that will please everyone, but we still have to try to coordinate everyone as well as possible.* »

- Diversify & engage

Also, a key recommendation is that algorithm design should be collegial. This can take different forms, but it has a common goal: the conceptor should never be alone in the conception process.

The most recurring recommendation is to introduce more diversity into algorithm design teams: more women, ethnic and social mix, age differences, etc. This would considerably reduce the subjectivity of algorithm conception, Mrs. Lair gives an example illustrating this need: « *There was a controversy around Apple's Health application, which was only made by men, and they forgot to include that women had their periods. If there had been only one woman on the team, it wouldn't have happened. Diversity in teams is very important.* »

Also, the current distance between those who think about ethical issues and those who develop algorithms must be reduced: they are often completely separate people who do not have a good knowledge of the other field. To overcome this, AI concepots should be trained and made aware of the ethical issues and bias in the systems they develop. More broadly, there is a need to educate the public about what an algorithm is and how an AI works. All stakeholders, especially users and conceptors, must be trained so that they are better aware of the services they handle and the risks associated with them: « *People must be familiarized with these notions so that they are aware of them and have the right knowledge. For example, a marketing department that orders an algorithm must be able to understand these subjects in order to read the algorithm and check that there is no discrimination or bias. Education is the first step.* » (Mrs. Lair)

Audit & verify

Some more operational solutions were also mentioned by respondents. These are concrete initiatives that can be easily implemented at the level of organizations.

- Check & adjust

Combating statistical bias is the most tangible approach: a statistical bias-free algorithm is one that tends to be more neutral.

A fundamental good practice to implement is data control: « *The higher the quality, completeness and diversity of the* data, *the less likely it is to be biased.* » (Dr. De La Roche) Conceptors need to have the best possible data, to adjust them where feasible and necessary, and to be very careful in making changes because this can generate other bias. A second point concerns the selection of the AI algorithm in the strict sense: conceptors need to carefully chose the right machine learning model, while being aware of its advantages and flaws. Dr. Boutalbi mentioned this avenue of resolution: « *We need to think about new models that are less subject to bias.* »

Also, Mrs. Lair recommend that « Responsible AI » type job be systematically created in companies using AI, on a full-time or part-time basis, in order to have guarantors of the ethics of

algorithms in organizations. After the algorithm has been created, and before it is put into production, mandatory validation processes must also be put in place: this is Dr. De La Roche's job called « Audit and Validation ». This consists in testing the algorithms, in real and diverse situations, to ensure their good reaction. The audit can notably go through quality processes, i.e., guides containing good practices, typologies of possible errors, precise control points, etc.

Clearly, algorithms must also be dealt with a « case-by-case » basis regarding ethical issues. For example, as Mr. Champion mentioned, the LAWS (Lethal Autonomous Weapon Systems), also known as « killer robots », can comply with the rules advocated by IHL (International Humanitarian Law). Even if the behaviour of the algorithm will never be perfect and will always be confronted with new and unlearned situations, IHL contains « simple » rules that are more or less easily integrated into algorithms.

- Keeping the hands on the wheel

From a general point of view, for all the people interviewed, the best safeguard against bias is human. The final decision taken by an AI, especially in critical situations, must be systematically validated by a human: « *Having a human in certain critical final decisions ensures that everything that could not be coded in the algorithms is filled by a decision in a hierarchical chain.* » (Mr. Champion) As Mr. Champion points out, the « man-in-the-loop », « man-on-the-loop » and « man-out-of-the-loop » approaches comes from the military and is reminiscent of LAWS' that cannot fire without human consent. This control can take place to a more or less direct degree. Putting this balance into place in decision making seems to be a good solution because humans counter the bias of algorithms and vice versa: « *Humans must retain final decision-making power. The algorithm must be able to advise, to be a decision aid, and that's where it works best: when the two complement each other.* » (Dr. Bardolle) Mrs. Lair also support this idea: « *It allows to make iterations to adjust the dysfunctions.* »

Moreover, this need to keep the human being in the decision-making process is linked to two notions.

The first, and most obvious, is that most algorithms nowadays do not achieve a high enough level of accuracy to be trusted « blindly ». Moreover, as we have seen, at the present time it is impossible to reach a 100% accuracy level, and that's probably never gonna be possible. Thus, if an algorithm makes a mistake in critical decision-making without a human in the loop, then the question of liability arises. For example, in a traditional car, if there is an accident, then it is the driver's responsibility at stake. On the other hand, if it is in a completely autonomous car, then the responsibility lies on the manufacturer. As a result, manufacturers protect themselves by selling, for the moment, only « semi-autonomous » vehicles: the driver must remain attentive with his hands on the wheel. Furthermore, this is also a legal obligation, as States are vigilant with regard to the limits of these technologies. For Dr. De La Roche: « *The driver is always responsible. He must keep his hands on the steering wheel, keep an eye on things and take over in the event of a problem to avoid accidents. This is how car manufacturers protect themselves, but French legislation also prohibits the person behind the wheel from sleeping, for example. The present state of the machine learning and the current capabilities of vehicles are not sufficient to give total autonomy: today, yes, humans must be able to keep their hands on the wheel.* »

The second reason is that algorithms need to gain public trust, and this requires the presence of humans in the decision-making process. These newer technologies are still little-known to researchers and even more so to the public, who may be afraid of them. For example, in the

case of a diagnostic and medical recommendation algorithm, it is more psychologically reassuring to have a doctor to confirm the suggested treatment, even if the algorithm is more effective than the doctor. Mrs. Lair explains this confidence mechanism: « *This transition with humans is necessary to verify that we are not counterproductive and that we do not generate unintended consequences. Moreover, the notion of trust is crucial, which is why the European Commission's report is called Trustworthy AI. In order for people to accept that decisions are made by machines, there must be trust, and this requires the presence of humans in the loop.* »

Institutional players

There may be pressures from outside the organizations that could lead to changes in favour of combating bias and ethical practices. The introduction of attractive regulations or certifications could encourage companies to incorporate certain values into their services.

- Enforcement measures

From a legal point of view, some recommend coercive measures by the authorities, in order to force private actors to respect certain rules such as explicability, transparency, etc. Legislation can be a promising avenue since it would put the stakes of bias and ethics back at the heart of the interests of private companies because, as we have mentioned, disinterest in these issues creates situations conducive to their emergence. In this sense, Dr. Masure makes an analogy with the issues of personal data and privacy: « *We notice that people trust Apple and Google more than the State to control privacy issues. One reason for this is that in the event of a scandal, there is a huge economic stake for companies.* »

- Incentives measures

Keeping in mind that legislation is not always up to date on technological issues and cannot cover all situations specifically, voluntary approaches can also be effective. Competent bodies such as the IEEE can assist companies in their ethical approaches. In addition, institutions such as the European Commission have created *Trustworthy AI,* which is a set of recommendations and best practices related to artificial intelligence. Also, Mrs. Lair evokes this new willingness on the part of some companies to develop responsible technologies: « *Today, we even talk about Technological Social Responsibility (TSR) […], in addition to Corporate Social Responsibility (CSR) initiatives.* » In the same spirit, the Data For Good organization created the « *Hippocratic Oath for Data Scientist* » to empower data artisans.

However, there are limits to these practices: « *It is not enough to make people individually responsible, the system must also be responsible. If there is only a checklist and everyone individually respects it, it is also a way for the big players to make sure that they will not be regulated by the legislative bodies.* » (Dr. Bardolle)

A final element is the funding of research by public institutions, in order to have the tools to debias and respect algorithmic fairness. It is also necessary to federate researchers because the issues are complex and ambitious. According to Mrs. Lair, GAFAM could quickly have a monopoly on these tools because they are already working and progressing on these subjects.

Moreover, now that the world is aware of bias, it seems that they are the main limit to the use of algorithms. Then, if we manage to debias artificial intelligences « *by freeing them from biased human decisions* » (Mrs. Lair), we will be able to reach their immense potential, meaning using them on a large scale for the common good, without risking harm to anyone.

Recalibrate expectations

Beyond the tracks mentioned, some people take a very different critical point of view, calling for a strong stand back on the abilities of algorithms and the connection we have with them.

- Relativization

For the majority of those interviewed, the stronger the impact of the algorithm on humans, the more critical the ethics in algorithms becomes: justice, health, education, access to employment, credit scoring, etc., are sectors that cannot do without these issues. For Dr. De La Roche: « *If it doesn't influence human lives, it's not useful, but the more it influences them, the more critical it becomes.* » For example, a text character recognition algorithm has no ethical impact at first glance. This is in line with the « case-by-case » approach, arguing that each algorithm, depending on its impact and sector, has its own issues to deal with individually. As Mrs. Lair points out: « *It is in the fields of recruitment or justice that algorithms first caused controversy, whereas they have been present for longer in industry or construction.* »

Also, as Dr. De La Roche recommends, we must above all try to do better than humans, rather than trying to find the perfect accuracy. Overtaking humans is possible in many situations. For example, it seems impossible to create an autonomous car that will cause absolutely no accidents, or that will always make the right decision in a dilemma. Nevertheless, autonomous cars nowadays are already much safer and more responsive than a human driver.

- Emancipation is not yet a reality

It should be remembered that these technologies are recent and are constantly improving. Research today tends to make systems increasingly autonomous, to the point where they no longer require human intervention.

However, these systems are still far from being independent: « *Today, we cannot do without human beings to train an algorithm. You can get it to do a task very well, but it is nothing without humans, because they are the ones who give it the data on which it will learn. Most of the work that needs to be done is on the human side.* » (Dr. Boutalbi) Indeed, most of the data that make up the learning dataset are emitted by humans, through platforms such as Amazon Mechanical Turk. These services employ individuals paid per click to perform repetitive and tedious tasks, such as recognizing an object in a video or classifying synonyms of words.

Moreover, the importance of human judgment in the upstream processing of data demonstrates, among other things, the low autonomy of machine learning algorithms. Learning is a meticulous and time-consuming work, which can easily lead to errors (Broussard, 2018). For Dr. Bardolle: « *All these elements show that we are still far from the fantasy vision of the future with robots talking to us. My two-year-old son can do infinitely more than a machine driven by millions of dollars and billions of gigabytes of data.* » In addition, Mr. Pineau says that the machine's removal of human bias is double-edged: « *This is the example of American judges who, in an Israeli study, were found to be more lenient after lunch, when they were full, than before, when they were hungry and tended to rush the cases they had to judge. Will an algorithm that replaces judges' decisions be less biased? Probably not. Schedule bias will be eliminated because it is very human, but other forms of bias would be found in the data.* »

Indeed, as Dr. Bardolle points out, algorithms were invented to solve small calculations and small operations, not to process big data and act autonomously. The algorithms are so simply

biased because they are not used in a way that is consistent with their nature and usefulness: « *Machine learning algorithms work well when they have a fairly specific task to perform. On the other hand, all edge cases, i.e., complex cases, are difficult to handle. When it comes to differentiating a cat from a dog with a corresponding dataset, there is no problem, but when you show a fox, then it is confused. [...] Algorithms are not basically made to solve bias, they are made to solve specific and narrow problems, otherwise they will simply reproduce the bias present in the data.* »

These elements make us wonder about the use we make of algorithms. Are we expecting too much from them? Is their use in line with their capabilities? « *Machines are efficient for specific tasks, but as soon as you go out of their scope, everything collapses*. » (Dr. Bardolle)

CONCLUSION

This thesis raised several issues revolving around the question of algorithms. The objective was to know the different sources of bias in algorithms, to discuss the different remedies, and finally to identify the different ethical challenges related to these issues.

There is no absolute classification of bias, as this can vary depending on the approach and the level of granularity of the perspective adopted. However, in order to be as exhaustive as possible, the approach chosen is « chronological », allowing the bias to be classified according to their order of appearance in the life cycle of the algorithm: data, design, use. When the first category is often shown to be guilty, the other two tend to be forgotten.

From the literature review and the research conducted, it is clear that data bias is without a doubt the bias with the greatest impact and the most noticed. This is because data is the heart of algorithms and generally produces the most glaring errors. Data bias is most often explained either by the fact that they are empirical, or by the psychological bias of the conceptors who manipulate them: this is true, but other bias can also appear. The statistical processing of data, their quality, completeness, the passage of time, or the use of reinforcement learning are also contributing factors.

The algorithmic bias most evoked by the experts is the one relating to a bad choice of machine learning model, or relating to the subjective choices of conception. There are also other sources of bias, linked to the limits of the algorithm condition or, more broadly, to the limits of statistics. The fact that the algorithm learns from past data, that it is completely dependent on its data, and that it incorporates only elements that are quantifiable, or simple enough to be quantified, are also determining elements.

Usage bias, since they are extrinsic qualities of the algorithm, are often overlooked by people who are not designers. Yet, the way in which the results are formalized, the interfaces, and the context of their use are inseparable from the algorithms. Thus, a lot of bias appears in the use, a level particularly prey to the subjectivity, interpretation and influence of the users interacting with the system.

Also, big data is often mistakenly considered as a miracle solution to algorithmic bias. This is because an extremely large volume of data generates a sense of completeness and therefore overconfidence among conceptors. Still, a dataset that is considered to be « big data » is particularly prone to bias, especially because of its size, and quantity is not a silver bullet.

Consequently, it is important to remember that omission, technopositivism and presumption, i.e., all the elements leading to a decrease in vigilance regarding bias, create situations that are particularly conducive to their emergence. Indeed, since bias are inherent to algorithmic technologies, the first necessity is to identify them, because the very nature of bias is that it manifests itself without our knowledge. It is therefore essential that conceptors and stakeholders have the best knowledge possible of the potential sources of bias, from data to use, through all the intermediate elements of the chain, at all levels of granularity.

Beyond the identification stage, it is legitimate to seek to correct bias, either to increase efficiency or to avoid errors and harmful consequences.

Different operational methods are available to data scientists: statistical methods, documentation, use of debiasing tools and frameworks, etc. However, these are approaches that can be described as corrective, i.e., they correct the bias that have manifested themselves a posteriori, but they do not address the source of the problem. In addition, some recommend the use of transparent or « auditable » algorithms, i.e., algorithms intelligible allowing to be examined to identify potential bias, sometimes on a public scale. However, it is not possible to make all algorithms public, either for business reasons, including intellectual property, or for legal reasons, including privacy. In addition, machine learning usually generates black boxes, whose opacity is positively correlated to the complexity of the algorithm, and as algorithms become more and more complex, one can expect a multiplication of black boxes. The problem with black boxes is that they make it particularly difficult to understand the algorithms and, indeed, to identify bias and their causes. As a result, it is like trying to square the circle: the more complex an algorithm is, the more bias it is subject to, but the less intelligible it is. These elements contribute to increase the fear around artificial intelligence technologies and to degrade the public's trust in them.

While algorithmic bias can have consequences that simply produce statistical errors, they can also produce ethical issues. In critical situations such as justice or health, for example, algorithmic bias can arise serious consequences, especially since they amplify and automate the bias. To overcome this, various methods are being considered.

Some advocate methods that attack the « source » of bias: reducing societal bias would directly reduce bias embedded in the data and thus debias the algorithms. For example, if recruitment algorithms are gender biased, it is because they learned from recruitment data that were gender biased. Thus, if companies were less sexist in recruitment, the algorithm would be less sexist as well. The problem with this approach is that it's very long and intangible, it could take decades. Thus, some researchers advocate a more effective approach, which is to debias the algorithms first, which will help to debias the society retroactively, rather than the other way around.

Moreover, explainable AI is a track that is gaining in popularity because it is obviously impossible to make ethical a reasoning that is not understandable, withal, this is incompatible with the development of complex algorithms. Also, some researchers have developed operational methods to solve ethical dilemmas: statistical formalization of ethical rules, frameworks, public platforms for collecting the public opinion, etc. However, ethics is a plural notion: it is subjective and it is sometimes impossible to reach statistical compromises that satisfy everyone and meet all ethical approaches and requirements. Moreover, in many situations, it is impossible to predict all the use cases that an algorithm will encounter, so algorithms will always come up against new cases, generating errors and confusion.

Also, algorithmic ethics is a rising concern, becoming a business issue for companies, and the pressure of public opinion plays a key role in this. However, the application of ethics in algorithms is sometimes incompatible with the business model of companies. Also, it requires time-consuming and costly steps, with no guarantee of results. In addition, there is the emerging issue of liability in case of prejudices, which is diluted with the emergence of black boxes, while the damages are very real: companies are afraid of legislative sanctions that will become tougher over time. These conjunctions are leading companies towards « ethics washing », i.e., the use of fallacious methods to whitewash their image and to promote the use of a front, rather than a background, ethical approach.

It should also be noted that computer science, ethics and interface design are complex fields of expertise. These are compartmentalized and partitioned areas whose knowledge is mastered by distinct individuals who do not spontaneously collaborate with each other. Yet, if the best possible algorithmic equity is to be achieved, it is imperative to coordinate these fields of expertise. As such, it is important to develop algorithmic issues in a collegial way: it is necessary to decide commonly on the ethical rules to be put in place and to be aware of the dangers represented by the massive diffusion of an algorithm. The creation of multidisciplinary and independent committees should be encouraged, in the image of what is done in the field of bioethics: population, philosophers, developers, politicians, companies, etc. All stakeholders must be involved in answering these questions, on the one hand to reduce the more obvious bias and, on the other hand, to « encode » algorithmic ethics that is appropriate to suit as many people as possible.

In addition, it is also necessary to diversify the conceptions teams: more diversity will automatically reduce discriminatory bias. Therefore, companies must adapt by creating *Responsible AI* positions within their organization, by implementing rigorous processes of control and validation to fight bias and by encouraging the training of its conceptors in ethical issues.

Finally, the balance between human and machine seems, given the current state of machine learning, to be a key issue in critical impact algorithms. Humans need to be able to control and counteract the emergence of algorithmic bias, and algorithms need to be able to inform about the emergence of human bias. The relationship between humans and algorithms in the decision-making process seems, for the moment, to be the best way to create synergy and to avoid bias being present at the time of production and spreading.

Moreover, since the emergence of bias is almost systematic in algorithms, some people think that it is impossible to obtain algorithms free of all bias. Consequently, this leads to questions about the use we make of this tool, raising two issues. First, some people think that we have too high expectations among algorithms and that debiasing them is a vain goal: they « simply » have to do better than humans. On the one hand, this goal is achievable in many situations, and on the other hand, it already provides significant benefits. Secondly, some people question the use we make of algorithms: they are good for narrow and determined tasks, but are extremely sensitive to bias as soon as they become more sophisticated. They are limited instruments, doomed to generate bias when used for complex activities, and much less autonomous than common thinking would suggest.

We therefore recommend vigilance as well as the systematic adoption of a critical point of view during the conception stage of algorithms and also afterwards. Diversity and collegiality should be promoted in algorithm development and in ethical thinking, both within and outside organizations, including seeking to align the interests of companies with those of their users. It is necessary to set achievable goals and to get rid of the thought that a neutral algorithm is conceivable: we must seek to reduce as much bias and to have the most ethical impact as possible, while keeping in mind that results will never be absolute. For the time being, algorithms with critical impacts cannot do without humans and we must also seek to inform users about the reality of the tools they use on a daily basis.

This thesis also has its limits. We stress the fact that we took a general view on the stake of algorithms, but that each use and each sector has its own particularities. The statistical and

human issues may be very different from one situation to another, and more in-depth and case-specific research is needed.

Beyond the question of algorithms, we have touched on other problems related to the rise of digital technologies in our environment. The more these technologies evolve and impact our world, the more we discover their flaws, weaknesses and vices. Generally speaking, we should seek to align, harmonize and balance, once again, business interests with those of users. But let there be no mistake: while some emerging issues arise from asymmetry of interests, others are unintentional and also represent a danger. Also, the issues related to captology are interesting researches for the future.

REFERENCES

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7), 16-07.

Baeza-Yates, R. A. (2013). Big data or right data?. In AMW.

Baeza-Yates, R. A. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54-61.

Baum, S. D. (2017). Social choice ethics in artificial intelligence. AI & SOCIETY, 1-12. DOI 10.1007/s00146-017-0760-1.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Bertail, P., Bounie, D., Clémençon, S., & Waelbroeck, P. (2019). Algorithmes: biais, discrimination et équité.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, 277-284.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and information technology*, *15*(3), 209-227.

Broussard, M. (2018). Artificial unintelligence: How computers misunderstand the world. MIT Press.

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2019). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?. *arXiv preprint arXiv:1912.07398*.

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2018). Motivations and risks of machine ethics. Proceedings of the IEEE, 107(3), 562-574.

Courtland, R. (2018). Bias detectives: the researchers striving to make algorithms fair. *Nature*, *558*(7710), 357-357.

Daelemans, W., Van Den Bosch, A., & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. Machine learning, 34(1-3), 11-41.

Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. In *IJCAI* (pp. 4691-4697).

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

Eustache, S., & Trochet, J. (2017). De l'information au piège à clics. *Le Monde Diplomatique*. https://www.monde-diplomatique.fr/2017/08/EUSTACHE/57804

Favaretto, M., De Clercq, E., & Elger, B. S. (2019). Big Data and discrimination: perils, promises and solutions. A systematic review. Journal of Big Data, 6(1), 12.

Floridi, L. (2019). Translating principles into practices of digital ethics: five risks of being unethical. Philosophy & Technology, 32(2), 185-193.

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

Fuchs, D. J. (2018). The dangers of human-like bias in machine-learning algorithms. Missouri S&T's Peer to Peer, 2(1), 1.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential bias in machine learning algorithms using electronic health record data. *JAMA internal medicine*, *178*(11), 1544-1547.

Goel, S., Broder, A., Gabrilovich, E., & Pang, B. (2010, February). Anatomy of the long tail: ordinary people with extraordinary tastes. In Proceedings of the third ACM international conference on Web search and data mining (pp. 201-210).

Greene, D., Hoffmann, A. L., & Stark, L. (2019, January). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Hoste, V., & Daelemans, W. (2005). Comparing learning approaches to coreference resolution: there is more to it than 'bias'. In *Workshop on Meta-Learning; held in conjunction with the 22nd International conference on Machine Learning (ICML 2005)* (pp. 20-27).

Howard, A., Zhang, C., & Horvitz, E. (2017, March). Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In 2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO) (pp. 1-7). IEEE.

Johnson, I., McMahon, C., Schöning, J., & Hecht, B. (2017, May). The Effect of Population and" Structural" Bias on Social Media-based Algorithms: A Case Study in Geolocation Inference Across the Urban-Rural Spectrum. In *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems* (pp. 1167-1178).

Karapapa, S., & Borghi, M. (2015). Search engine liability for autocomplete suggestions: personality, privacy and the power of the algorithm. *International Journal of Law and Information Technology*, *23*(3), 261-289.

Leavy, S. (2018, May). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering* (pp. 14-16).

McFarland, D. A., & McFarland, H. R. (2015). Big data and the danger of being precisely inaccurate. *Big Data & Society*, *2*(2), 2053951715602495.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

Mooney, R. J. (1996). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. *arXiv preprint cmp-lg/9612001*.

Naughton, J. (2019). To err is human – is that why we fear machines that can be made to err less?. *The Guardian*. https://www.theguardian.com/commentisfree/2019/dec/14/err-is-human-why-fear-machines-made-to-err-less-algorithmic-bias

O'neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

Osoba, O. A., & Welser IV, W. (2017). An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation.

Pitoura, E., Tsaparas, P., Flouris, G., Fundulaki, I., Papadakos, P., Abiteboul, S., & Weikum, G. (2018). On measuring bias in online information. *ACM SIGMOD Record*, *46*(4), 16-21.

Price, M., & Ball, P. (2014). Big data, selection bias, and the statistical patterns of mortality in conflict. *SAIS Review of International Affairs*, *34*(1), 9-20.

Raub, M. (2018). Bots, bias and big data: artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. Ark. L. Rev., 71, 529.

Roselli, D., Matthews, J., & Talagala, N. (2019, May). Managing Bias in AI. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 539-544).

Rudman, L. A. (2004). Social justice in our minds, homes, and society: The nature, causes, and consequences of implicit bias. Social Justice Research, 17(2), 129-142.

Seely-Gant, K., & Frehill, L. M. (2015). Exploring bias and error in Big Data research. *Journal of the Washington Academy of Sciences*, *101*(3), 29-38.

Segal, D. (2011). The dirty little secrets of search. *The New York Times*, *12*(02). https://www.nytimes.com/2011/02/13/business/13search.html

Shadowen, N. (2019). Ethics and bias in machine learning: A technical study of what makes us "good". In *The Transhumanism Handbook* (pp. 247-261). Springer, Cham.

Shepperd, M., Bowes, D., & Hall, T. (2014). Researcher bias: The use of machine learning in software defect prediction. IEEE Transactions on Software Engineering, 40(6), 603-616.

Tam, S. M., & Kim, J. K. (2018). Big Data ethics and selection-bias: An official statistician's perspective. Statistical Journal of the IAOS, 34(4), 577-588.

Thompson, C. J. (2019). The 'big data' myth and the pitfalls of 'thick data' opportunism: on the need for a different ontology of markets and consumption. J*ournal of Marketing Management*, 35(3-4), 207-230.

Thornton, S. M., Pan, S., Erlien, S. M., & Gerdes, J. C. (2016). Incorporating ethical considerations into automated vehicle control. IEEE Transactions on Intelligent Transportation Systems, 18(6), 1429-1439.

Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, *5*(3), 246-255.

Wu, S., Rizoiu, M. A., & Xie, L. (2019). Estimating attention flow in online video networks. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-25.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. *arXiv preprint arXiv:1812.02953*.

Yue, Y., Patel, R., & Roehrig, H. (2010, April). Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th international conference on World wide web* (pp. 1011-1018).

Zeller Jr, T. (2006). A New Campaign Tactic: Manipulating Google Data. The New York Times, 26. https://www.nytimes.com/2006/10/26/us/politics/26googlebomb.html

APPENDIXES

I. Interview guide

Introduction

1. Pouvez-vous présenter votre parcours ?
   Can you introduce your career path ?

Biais & résolutions

2. Cette première partie est centrée sur les biais algorithmiques. Comment définissez-vous un algorithme biaisé ?
   This first part focuses on algorithmic bias. How do you define a biased algorithm?

3. Quelles sont les sources de biais dans les algorithmes ?
   What are the sources of bias in algorithms?

4. Quels sont les risques liés à une absence de traitement des biais dans les algorithmes ?
   What are the risks associated with a lack of treatment of bias in algorithms?

5. Quels sont les biais les plus difficiles à traiter dans les algorithmes ?
   What are the most difficult bias to deal with in algorithms?

6. Selon vous, quelles sont les meilleures approches pour résoudre les biais dans les algorithmes ?
   In your opinion, what are the best approaches to solve bias in algorithms?

7. Au sein de votre expérience professionnelle, avez-vous un exemple d'une mesure que vous avez mise en place pour résoudre un biais algorithmique ?
   In your professional experience, do you have an example of a measure that you have implemented to solve algorithmic bias?

Éthique & application

8. Nous arrivons à la deuxième partie des questions, davantage centrée sur l'éthique dans les algorithmes. Comment définissez-vous l'éthique ?
   We come to the second part of the questions, which is focused on ethics in algorithms. How do you define ethics?

9. Quelles sont les difficultés liées à l'application de l'éthique dans les algorithmes ?
   What are the difficulties related to the application of ethics in algorithms?

10. Selon vous, l'éthique peut-elle être objectivement appliquée dans les algorithmes ?
    In your opinion, can ethics be objectively applied in algorithms?

11. Y a-t-il des domaines où l'éthique n'est pas importante dans les algorithmes ?
    Are there areas where ethics is not important in algorithms?

12. Selon vous, qui doit décider de l'éthique à mettre en place dans les systèmes algorithmiques ?
    In your opinion, who should decide on the ethics to be implemented in algorithmic systems?

13. Au sein de votre expérience professionnelle, avez-vous un exemple d'un challenge éthique lié aux algorithmes auquel vous avez fait face ?
    In your professional experience, do you have an example of an ethical challenge related to algorithms that you have faced?

Conclusion & ouverture

14. Nous arrivons aux questions finales d'ouverture. On parle souvent de l'équilibre entre humains et machines à mettre en place dans le processus de décision, quel est votre point de vue ?
    We come to the final opening questions. We often heard about the balance between humans and machines in the decision-making process, what is your point of view?

15. Quels sont les grands enjeux actuels ou à venir de la recherche concernant l'éthique des algorithmes ?
    What are the major current or future research issues concerning the ethics of algorithms?

16. Quel est votre point de vue sur la place des biais dans les algorithmes pour la société à long terme ?
    What is your point of view on the place of bias in algorithms for society in the long term?

17. Nous avons terminé. Souhaitez-vous revenir sur l'un des sujets évoqués, préciser un élément ou poser une question ?
    We are done. Would you like to come back to any of the topics raised, add anything, or ask a question?

II. Interview transcripts

#1. DR. BARDOLLE

Date	Length	Communication
May 13, 2020	27′ 13″	Zoom - Video call

Pouvez-vous présenter votre parcours ?

Je m'appelle Frédéric Bardolle. À l'heure actuelle je suis le chef de la Fabrique Numérique, un incubateur de services numériques appartenant au Ministère des Armées. Mon travail sur l'éthique est issu essentiellement du monde associatif.

J'ai été l'un des responsables de l'association Data For Good pendant plusieurs années, ayant pour mission de réaliser des projets de data sciences pour l'intérêt général. Dans ce cadre, j'ai participé notamment à deux projets liés à l'éthique.

Le premier est Algotransparency, qui démontrait comment les algorithmes déterminant l'accès à l'information peuvent être biaisés et encourager la diffusion de fausses informations.

Le deuxième projet se nomme Le Serment d'Hippocrate des Data Scientists, qui consistait à mettre en place une check-list pour rendre plus responsables les gens travaillant et utilisant les données.

Cette première partie est centrée sur les biais algorithmiques. Comment définissez-vous un algorithme biaisé ?

Je vais faire le parallèle avec les humains : quand un humain a un biais, c'est souvent le fait qu'il a un stéréotype ancré en lui sans le savoir ; c'est une pensée spontanée, face à une personne ou à un évènement.

Pour un algorithme, c'est lorsqu'il réagit de manière différente selon certaines catégories de personnes ou d'entités, sans que cela soit voulu ou souhaitable par les concepteurs. Par exemple, il y a encore quelques mois, l'algorithme de Google Translate traduisait différemment le genre de certains mots d'une langue neutre, comme l'anglais, vers une langue à genre, comme le français ou l'espagnol. Les mots « doctor » et « nurse » étaient traduits en « docteur » ou « infirmière ». Ceci est un genre typique de comportement non souhaitable, dans le sens ou cela entretien des stéréotypes, ici liés au genre. Il y a donc un travail à faire de la part des concepteurs pour résoudre ce genre de problème. Google a d'ailleurs mis en production, il y a quelques semaines, une version corrigée de son algorithme proposant une traduction affichant désormais les deux genres lorsque c'est nécessaire.

Quelles sont les sources de biais dans les algorithmes ?

Par essence, les algorithmes d'apprentissage automatique sont conservateurs car ils réutilisent les données du passé. Pour Google Translate, l'algorithme est entraîné sur des mêmes ouvrages dans des langues différentes, puis, quand il y a suffisamment de données, il est mis en production et capable de reconnaître les schémas appris. Cependant, comme les données proviennent du passé, notamment car les quantités nécessaires sont importantes et qu'il faut

donc remonter dans le temps, on va observer des comportements non souhaitables. Ici, on reproduit les biais humains du passé. Si un algorithme traduit « doctor » au masculin, c'est parce qu'il l'a vu dans beaucoup de textes, et que les sociétés humaines étaient comme ça, mais comme les sociétés changent, les algorithmes n'ont pas le temps de s'adapter.

Un autre problème peut exister, et il concerne notamment le cas des recommandations de vidéos YouTube. Lorsque l'on cherche un terme, YouTube valorise les contenus les plus controversés et polémiques, souvent liés à des théories du complot par exemple. Pourquoi l'algorithme fait-il cela ? Il agit comme ça car il fait très bien son travail, et ce travail est simple : les algorithmes sont optimisés avec des objectifs, et ces objectifs sont une notion clé. L'objectif de l'algorithme de YouTube est de faire en sorte que les gens passent le plus de temps possible sur la plateforme. S'apercevant que les contenus complotistes font rester les gens plus longtemps, il va alors les diffuser massivement et renforcer ses contenus. Il y a d'ailleurs une métaphore assez amusante à ce sujet : c'est comme si dans une cantine pour enfants, le seul objectif d'un algorithme faisant les menus était « *Le plateau est-il vide ou non ?* ». Comme c'est automatisé, il ne peut que peser les plateaux pour voir s'ils sont vides ou pas. En testant de nombreux plats différents, il s'apercevra à la fin, que, en proposant des nuggets, des frites ou des glaces, les plateaux sont finis. Donc au bout de quelques jours les enfants ne mangeront que des choses très grasses et très sucrées. L'objectif que l'on mesure est-il le bon ? Faut-il que les enfants finissent leur plateau ? Non, le bon objectif est de nourrir des enfants et qu'ils soient en bonne santé. Cependant, c'est compliqué pour un algorithme, car il faudrait échanger avec les enfants en amont, tester sur le plus long terme, etc. C'est pareil pour YouTube : pour corriger ce biais, il faudrait se demander ce qui, en tant qu'être humain, nous apporte de la valeur, et ce n'est probablement pas de passer une nuit à regarder des vidéos nous affirmant que la terre est plate. Je pense que c'est plutôt d'apprendre des choses, se cultiver, écouter des musiques que l'on aime, etc. Mais ça, c'est beaucoup plus dur à mesurer que le temps que j'ai passé sur la plateforme.

Quels sont les risques liés à une absence de traitement des biais dans les algorithmes ?

Dans le cas des algorithmes de recommandation, on risque d'assister a une polarisation de plus en plus grande des idées. Plus on va avancer dans le temps, plus les gens vont proposer des idées extrêmes. D'ailleurs, ils n'y croient pas forcément eux-mêmes, mais leurs vidéos sont vues et ça leur rapporte de l'argent. Je pense que beaucoup de créateurs de contenu font ça actuellement, c'est-à-dire qu'ils parlent de sujets complotistes car les algorithmes les favorisent. S'ils parlaient de sujets scientifiques par exemple, ils auraient moins de vues. Non seulement l'algorithme diffuse un certain contenu, mais en plus il a un effet incitatif à créer ce genre de contenu : c'est un grand risque.

Pourtant, à l'origine, c'était une chance. La promesse de l'algorithme de recommandation était incroyable : chaque personne ayant des passions, même de niche, peux avoir accès au contenu qui lui plaît. En revanche, sur des sujets comme la politique par exemple, le risque est tout autre. Qui plus est, pour la science, il n'y a pas d'opinions, même si l'on peut challenger certaines idées et que le paradigme scientifique évolue constamment. Placer sur un pied d'égalité des opinions et la réalité scientifique est dangereux, car on risque de perdre ce que la science peut nous apporter via la méthode scientifique.

Quels sont les biais les plus difficiles à traiter dans les algorithmes ?

Tous ces biais sont difficiles à traiter. Les algorithmes d'apprentissage automatiques fonctionnent bien lorsqu'ils ont une tâche assez précise à effectuer. En revanche, tous les *edge cases*, c'est-à-dire les cas complexes, sont difficiles à traiter. Quand il s'agit de différencier un chat d'un chien avec un dataset correspondant, il n'y a pas de problème, mais quand on lui montre un renard, alors il est bloqué. Dans les exemples précédents, c'est la même idée à très grande échelle : si ce n'est pas pensé dès la conception, on arrive à ce genre de situation. Les algorithmes ne sont pas fait à la base pour résoudre des biais mais pour résoudre des problèmes précis et étroits, sinon ils reproduiront simplement les biais présents dans les données. Le fait de les corriger est un investissement énorme et incertain sur les résultats. Pour son problème de traduction, Google a dû créer un conseil composé de scientifiques, d'ingénieurs et de chercheurs, travailler pendant plusieurs mois et investir des sommes énormes, et ce n'est même pas corriger à 100%. Plus l'IA va se démocratiser, plus ce type de problèmes va se répandre. C'est difficile à traiter, et parfois on ne sait même pas faire. Par exemple, concernant la modération de contenu sur Facebook, c'est généralement signalé par des humains puis revu par des modérateurs. Ils passent des heures à traiter des contenus très violents, leur provoquant des dommages psychologiques important, mais actuellement ce n'est pas faisable par une IA et on ne sait pas éliminer ce genre de choses.

Selon vous, quelles sont les meilleures approches pour résoudre les biais dans les algorithmes ?

Au sein de Data For Good, on a travaillé sur le Serment d'Hippocrate des Data Scientists, sur ces questions qui intéressent tous les métiers liés à la data. Malheureusement, ça ne suffit pas de responsabiliser les gens de manière individuelle, le système aussi doit être responsable. S'il n'y a qu'une check-list et que chacun individuellement la respecte, c'est aussi une manière pour les grands acteurs de s'assurer qu'ils ne seront pas régulés par les organes législatifs.

Aussi, on avait fait une proposition à YouTube. On avait constaté qu'il y avait un problème entre le statut d'hébergeur web, qui n'a aucune responsabilité, et celui de média, ou il y a un responsable éditorial qui porte toute la responsabilité. Il n'y a pas de statut intermédiaire. Il est vrai qu'au début, un hébergeur était juste une plateforme stockant des vidéos avec un annuaire pour les répertorier, il n'y avait pas vraiment de responsabilité. Cependant, désormais les algorithmes « poussent » du contenu, c'est différent. Il faudrait envisager une responsabilité de la plateforme, non pas lorsqu'elle héberge du contenu, mais lorsqu'elle le recommande massivement. Par exemple, pour une vidéo recommandée un million de fois par un algorithme, n'y a-t-il pas une forme d'éditorialisation de la plateforme ? De fait, elle devient responsable de ce contenu, même si c'est fait par un algorithme, car c'est quand même géré par l'entreprise.

Au sein de votre expérience professionnelle, avez-vous un exemple d'une mesure que vous avez mise en place pour résoudre un biais algorithmique ?

Dans mon travail je n'utilise pas directement des algorithmes. Chez Data For Good, c'était principalement les projets que je vous ai évoqués.

Nous arrivons à la deuxième partie des questions, davantage centrée sur l'éthique dans les algorithmes. Comment définissez-vous l'éthique ?

L'éthique, selon moi, consiste à se demander quel est l'objectif de la vie humaine, individuellement et collectivement. Une fois l'objectif fixé, c'est aussi se demander comment faire pour y parvenir. Cette idée rejoint celle des objectifs des algorithmes : les objectifs que l'on se fixe en tant qu'individu ou que la société nous impose sont-ils les mêmes que ceux que l'on a assignés à nos algorithmes ? Le problème est que la réponse est souvent non.

Quelles sont les difficultés liées à l'application de l'éthique dans les algorithmes ?

Je pense que ça correspond principalement à cette notion d'objectif, à l'alignement entre ceux de notre société et ceux des algorithmes.

Selon vous, l'éthique peut-elle être objectivement appliquée dans les algorithmes ?

Une éthique par essence est subjective, ça n'est pas quelque chose de scientifique au sens de la méthode scientifique. Il n'y a pas d'éthique universellement définie ni universellement acceptée. L'éthique s'appuie sur des choses qui sont du domaine métaphysique, donc on ne peut pas rendre ça universel. Il y a des tentatives, comme la Déclaration des Droits de l'Homme, mais même ça, ça n'est pas accepté par tous les états.

Y a-t-il des domaines où l'éthique n'est pas importante dans les algorithmes ?

Le domaine de l'apprentissage automatique est très vaste : ça peut être reconnaître des images, des vidéos, des objets, etc. Il y a des domaines ou ça n'est pas nécessaire. Par exemple, pour un logiciel de reconnaissance de caractères sur papier pour ensuite le numériser, ça ne paraît pas important, il ne peut pas y avoir de biais éthiques. Si ça n'influence pas des vies humaines, ça n'est pas utile, mais plus ça les influence, plus ça devient critique.

Selon vous, qui doit décider de l'éthique à mettre en place dans les systèmes algorithmiques ?

Dans l'idéal, ça serait en collégialité entre ceux qui exploitent et ceux qui utilisent les algorithmes. Pour reprendre notre exemple, si on demandait aux utilisateurs si l'algorithme de YouTube fonctionne comme ils le veulent, la réponse serait souvent négative. On a tous déjà passé une nuit à regarder des vidéos car l'algorithme nous entraîne dans une spirale, et ça n'est pas ce que l'on voulait à la base. Il faut avoir des discussions pour mieux comprendre ce que veulent les gens. Il y a d'ailleurs des personnes aux États-Unis qui ont créé le Center For Human Technology autour de Tristan Harris. Ils se sont demandé comment notre temps, constituant notre principale richesse en tant qu'être humain, doit être dépensé de manière utile pour nous et non pour une entreprise.

Nous arrivons aux questions finales d'ouverture. Quel est votre point de vue sur la place des biais dans les algorithmes pour la société à long terme ?

Je pense qu'il y a un sujet qui n'est pas assez traité, c'est le fait qu'aujourd'hui, beaucoup de travail est fait par les humains et non pas par les machines. À ce titre, il y a les travaux très intéressants d'Antonio Casilli autour du concept de *digital labor*, qui montre que l'on automatise une partie des choses, mais qu'une grande partie des tâches pour améliorer les IA sont faites par des humains. La conséquence, c'est la naissance d'une nouvelle forme de délocalisation, où l'on envoie des flux de données à des gens payés au clique. Tous ces éléments montrent qu'on

est encore loin de la vision fantasmée du futur avec les robots qui nous parlent. Mon fils de deux ans sait faire infiniment plus de choses qu'une machine entraînée à coups de millions de dollars et de milliards de gigabyte de données, car il y a des choses qui ne sont pas à produire. Les machines sont efficaces pour des tâches précises, mais dès qu'on sort un peu de leur *scope*, tout s'effondre. D'ailleurs, peut être que c'est une bonne chose, car si l'on va encore plus loin, cela pourrait poser plus de problème que l'on en a déjà.

On parle souvent de l'équilibre entre humains et machines à mettre en place dans le processus de décision, quel est votre point de vue ?

Dès qu'il y a une vie en jeu, une influence sur une vie humaine, l'humain doit garder un pouvoir de décision final. L'algorithme doit pouvoir conseiller, être une aide à la décision, et c'est d'ailleurs là que ça marche le mieux, lorsque les deux se complètent.

#2. DR. BOUTALBI

Date	Length	Communication
May 14, 2020	27′ 55″	Zoom - Audio call

Pouvez-vous présenter votre parcours ?

J'ai obtenu un Master en Informatique en Algérie, puis je suis arrivée en France en 2015. Ensuite, j'ai fait un Master 2 en IA à l'Université Paris Descartes. Au cours de ce Master, j'ai réalisé un stage chez Véolia dans leur service de R&D à propos de la détection de contamination dans le réseau d'eau d'Île de France. À partir de là, j'ai décidé de faire une thèse en machine learning. J'ai donc poursuivi en réalisant avec une thèse CIFRE, c'est-à-dire qu'une entreprise finance notre recherche et nous répondons à l'une de ses problématiques. Cette entreprise est une start-up qui s'appelle Trinov et qui édite des logiciels pour la gestion des déchets. Mes travaux ont porté sur l'optimisation et l'émission de recommandations concernant la gestion des déchets. J'ai terminé cette thèse récemment et je suis désormais data scientist et chercheuse chez Trinov.

Cette première partie est centrée sur les biais algorithmiques. Comment définissez-vous un algorithme biaisé ?

Définissons d'abord ce qu'est un algorithme. Les algorithmes d'apprentissage permettent d'apprendre à une machine à reconnaître des objets dans des images, classifier des documents, etc. Il y a deux grandes familles de méthodes : les algorithmes d'apprentissage supervisés et ceux non supervisés. La différence est que, dans les approches supervisées, on connaît a priori les *labels*. Par exemple, pour un algorithme de reconnaissance de personne dans les images, on l'entraîne avec des images en lui précisant où est la personne. Dans ce cas, on sait comment l'algorithme raisonne pour retrouver les personnes dans les images. L'objectif, à terme, et de ne plus avoir à préciser en amont où se trouve la personne mais qu'il devienne capable de la retrouver quand même : ce sont les algorithmes d'apprentissage non supervisés. Pour eux, il n'y a pas d'a priori, pas de connaissance à propos de l'information cherchée : on va essayer de la découvrir.

Pour revenir à la question des biais, selon moi, un algorithme biaisé n'existe pas : l'algorithme est biaisé par les données d'entrée qui sont biaisées ; ce sont les informations contenues dans les données qui génèrent le biais. Un exemple tout simple est celui des chatbots utilisés par les entreprises sur leur site web à destination des internautes. Ces outils utilisent des algorithmes de machine learning. Ils ont été entrainés sur des milliers de conversations pour qu'ils puissent tenir une discussion en interaction avec le client. Dans ce cas là, si l'on entraîne le chatbot sur des conversations qui ne sont pas filtrées ou traitées en amont, on se retrouvera très probablement avec des biais raciaux, sexistes, etc. Il faut donc faire très attention à la façon dont on sélectionne le jeu de données et à la manière dont on test notre modèle d'apprentissage. Lorsque l'on créer ces modèles, il y d'abord une partie d'apprentissage où on l'entraine à réaliser une certaine tâche, puis il y a une partie de test où l'on soumet le modèle à des données nouvelles qui n'ont pas été utilisées dans l'apprentissage. Il faut être vigilant à ce que les données utilisées dans notre modèle soient équilibrées : il faut éviter notamment d'avoir des minorités. Par exemple, j'ai vu le cas d'un algorithme américain qui reconnaissait une arme à feu sur une image, mais une grande partie des images d'entrée avec des armes à feu étaient

détenues par des personnes de couleur noire. L'algorithme a donc été biaisé, et s'il reconnaissait une personne noire sur l'image, il donnait un plus grand poids à la probabilité qu'il y ait une arme à feu sur l'image. Ainsi, il faut faire très attentions aux données d'entraînement, et au fait que les données soient équilibrées.

Vous avez évoqué les données, identifiez-vous d'autres sources de biais dans les algorithmes ?

Chaque modèle d'apprentissage a ses spécificités, et il y en a des milliers. Il faut faire très attention au moment de sélectionner son modèle : certains sont plus sensibles que d'autres. Les modèles d'apprentissage non supervisés sont plus sujets aux biais que les modèles supervisés. Cela appuie la nécessité de bien tester son modèle avant de le mettre en production.

Quels sont les risques liés à une absence de traitement des biais dans les algorithmes ?

Il y a deux grand risques liés aux biais. Le premier risque est d'influencer une décision. Par exemple, c'est aux États-Unis que l'on trouve le plus d'algorithmes de machine learning mis en production dans la vie courante, c'est donc là où l'on remarque le plus de biais. Il y avait notamment le cas d'un algorithme pour évaluer les risques d'infractions lorsque les policiers faisaient leurs rondes. L'algorithme était entraîné sur des données qui intégraient la géolocalisation via les codes postaux. C'est très dangereux, car si l'on considère que le code postal entre en compte dans le fait qu'une personne soit à risque, ça n'a pas de sens. Ça peut notamment influencer les policiers dans la prise de mauvaises décisions.

Le deuxième risque est l'automatisation des biais : les biais deviennent alors une règle et cela peut causer des dommages à grande échelle.

Quels sont les biais les plus difficiles à traiter dans les algorithmes ?

Je pense que tout dépend de la problématique ou du modèle sur lesquels on travaille. En règle générale, je ne pense pas qu'il y ait de grande différence, ça dépend de la situation.

Au sein de votre expérience professionnelle, avez-vous un exemple d'une mesure que vous avez mise en place pour résoudre un biais algorithmique ?

Traiter un biais, c'est bien traiter ses données, c'est le plus important. Je travaille beaucoup sur la problématique de recommandation et d'optimisation, et pour créer un bon modèle d'optimisation, qui, par exemple, optimise au mieux les chemins de collecte de déchets, il faut les entraîner sur des données qui sont déjà optimales, sinon, les résultats ne seront pas optimaux : c'est une relation de cause à effet. Il faut également faire attention au choix de nos variables, comme nous l'avons évoqué avec le modèle de géolocalisation. Il vaut mieux supprimer une variable quitte à dégrader la qualité des résultats plutôt que de risquer d'obtenir ce genre de biais.

Nous arrivons à la deuxième partie des questions, davantage centrée sur l'éthique dans les algorithmes. Comment définissez-vous l'éthique ?

Avoir un comportement éthique selon moi, c'est de ne pas porter de préjudice à quelqu'un.

Quelles sont les difficultés liées à l'application de l'éthique dans les algorithmes ?

L'éthique n'est pas toujours considérée. Actuellement, il n'y a pas de loi qui condamne pour un algorithme biaisé, mais je pense que la réglementation va changer et la pression va s'accroître

envers les sociétés qui construisent ces systèmes. Cette problématique de biais est aujourd'hui très connue et doit être prise davantage en considération : il faut y porter plus d'attention.

Selon vous, l'éthique peut-elle être objectivement appliquée dans les algorithmes ?

Je ne pense pas car l'être humain en lui-même est imparfait et nous sommes biaisés dans notre vie au quotidien. Ainsi, je ne pense pas que l'on puisse avoir une éthique commune applicable.

Y a-t-il des domaines où l'éthique n'est pas importante dans les algorithmes ?

Oui, il y a des domaines qui sont moins critiques que d'autres. Dans la justice par exemple, les enjeux sont beaucoup plus importants que dans d'autres applications. Si notre modèle doit détecter des animaux sur une image, à première vue, je n'ai pas l'impression qu'il y ait d'enjeux éthiques : tout dépend de la situation.

Selon-vous, qui doit décider de l'éthique à mettre en place dans les algorithmes ?

Généralement, quand une entreprise met en place un système de machine learning, une équipe de data scientist est en charge mais c'est l'entreprise qui est responsable. De fait, elle doit mettre en place des process pour garantir cette éthique, car les data scientists sont en réalité des informaticiens qui peuvent omettre ou relativiser les risques de biais. Ce sont des êtres humains qui ont leur propre perception. Les entreprises doivent intégrer des méthodes mais aussi davantage de validation, de contrôle et de tests de ses modèles de machine learning.

Au sein de votre expérience professionnelle, avez-vous un exemple d'un challenge éthique lié aux algorithmes auquel vous avez fait face ?

Comme je vous le disais, je pense qu'il y a des domaine critiques concernant l'éthique et d'autres dans lesquels ça l'est beaucoup moins. Dans le domaine dans lequel je travaille, l'optimisation de la gestion des déchets, il y a surement des biais, mais il n'y a pas de problématique éthique majeur.

Nous arrivons aux questions finales d'ouverture. On parle souvent de l'équilibre entre humains et machines à mettre en place dans le processus de décision, quel est votre point de vue ?

Selon moi, l'humain doit rester dans la boucle, et c'est d'ailleurs une expression de plus en plus répandue pour souligner que l'Homme doit intervenir dans ces modèles. Aujourd'hui, on ne peut pas se passer d'êtres humains pour entrainer des algorithmes. On peut l'amener à réaliser très bien une tâche, mais elle n'est rien sans l'humain car c'est lui qui lui confère les données sur lesquelles elle va apprendre. Le plus gros du travail à faire est du côté humain, en étant plus vigilant sur ce que l'on donne à la machine, car il en va de notre responsabilité. Aussi, on devrait réfléchir à de nouveaux modèles moins sujets aux biais.

Quel est votre point de vue sur la place des biais dans les algorithmes pour la société à long terme ?

Je suis plutôt optimiste, car de plus en plus de grands chercheurs, tel que Yann Le Cun entre autres, en parlent, et c'est déjà un bon point. Ces problématiques sont de plus en plus intégrées dans les recherches donc c'est positif. On va devoir améliorer nos modèles et nos méthodes d'entrainement. Même si on ne pourra pas complètement éviter les biais, on pourra néanmoins les réduire.

#3. MR. CHAMPION

Date	Length	Communication
April 28, 2020	35′ 20″	Zoom - Audio call

Pouvez-vous présenter votre parcours ?

Je suis Renaud Champion et je suis ingénieur en intelligence artificielle. Cela fait une vingtaine d'années que je travaille dans ce secteur en tant qu'investisseur, entrepreneur, et aussi auprès de différents gouvernements et de grandes institutions telles que la Commission Européenne. Je travaille également avec des associations internationales comme l'IEEE (Institute of Electrical and Electronics Engineers) qui réfléchit justement aux enjeux éthiques, juridiques et sociaux économiques liées à l'IA et aux systèmes intelligents tel que les robots et les algorithmes.

Comment définissez-vous un algorithme biaisé ?

Si l'on parle d'un système d'IA, c'est une combinaison de différentes briques technologiques qui vont de la perception d'un environnement à l'analyse de données en passant par leur modélisation, afin de prendre une décision, de la concevoir, de l'apprendre et enfin de l'exécuter. Donc un système d'IA est quelque chose de complexe et il peut y avoir des biais qui interviennent à différents niveaux de ces systèmes.

Le premier biais, qui est le plus connu, est le biais de la donnée. C'est-à-dire que lorsque les capteurs du système d'intelligence artificielle vont capter l'environnement, ils vont prendre des données brutes et celles-ci peuvent être biaisées au sens où elles ne sont pas représentatives de la diversité de l'univers observé. Par exemple, si l'on regarde un robot qui se déplace dans la foule d'un aéroport, bien évidemment si ce robot est à Paris, en Asie ou dans un pays du Golfe, il ne va pas du tout remonter le même type de données, notamment le même type d'informations morphologiques des personnes qu'il va voir. De la même manière, un moteur d'IA qui analyse des commentaires sur des forums de discussion Twitter, en fonction de la liste des forums qui lui sont donnés en amont, les informations et les données brutes qu'il va récupérer ne vont pas être représentatifs de tous les forums et de toutes les opinions du monde sur un sujet donné. Ce sont uniquement les opinions des personnes présentes dans ce forum à l'instant où il y a une capture de de données qui sont représentées. Il y a donc un biais au niveau de la donnée parce que, potentiellement, il n'y a pas une représentation globale des phénomènes et de l'environnement que l'on veut observer à l'instant T : c'est le premier biais.

Le deuxième biais, que l'on appel biais algorithmique, c'est un biais au niveau de la modélisation mathématique que le moteur d'IA va mettre en place afin de prendre une décision. Il peut faire des arbitrages et émettre des aprioris dans la manière dont cette formule va être construite. Par exemple, l'entreprise Tinder a défini dans ses algorithmes un indice de désirabilité. Cet indice donne une note qui sera utilisée ensuite dans la recommandation que l'application va faire à ses clients. L'indice de désirabilité peut varier selon un homme ou une femme : cela signifie que la désirabilité pour une même caractéristique ne sera pas évaluée pareillement en fonction du sexe. Un exemple qui a été qui a été mis en avant est notamment celui de la situation professionnelle. De manière schématique, la situation professionnelle élevée pour un homme est jugée beaucoup plus désirable que pour une femme. En effet, l'indice de désirabilité entre un homme qui a une très bonne situation professionnelle et une

femme qui a une situation professionnelle équivalente vont être complètement inverse : c'est un biais algorithmique. On est donc vraiment à deux niveaux de biais, mais l'IA en tant que telle n'introduit pas des biais elle-même. Cependant, elle met un grand coup de projecteur sur les biais qui peuvent être introduits soit par les données, soit par les choix arbitraires que vont faire les humains dans la définition des caractéristiques de l'algorithme.

Quels sont les risques liés à une absence de traitement des biais dans les algorithmes ?

À ce sujet, je vous conseille vivement la lecture du livre de Cathy O'neil, *Weapons of Maths Destruction*, car il répond exactement à votre question. Elle travaille à partir d'études américaines sur les algorithmes tout en faisant une analyse personnelle en tant que chercheuse. Par exemple, dans le domaine de la justice où il y a des outils d'aide à la décision basé sur l'IA qui sont proposés aux juges américains, elle démontre que certains de ces algorithmes ont des biais raciaux ancrés produisant des conseils biaisés aux juges. Je vous laisse imaginer les conséquences sur les condamnations… Également, au niveau de la santé sur les questions de de pauvreté, une personne habitant dans un quartier identifié comme modeste a une probabilité plus faible de pouvoir payer une couverture santé. En conséquence, dans un hôpital donné, on pourrait lui refuser les soins. Il y a d'autres exemples sur la santé, la justice, etc. Il y a des vrais enjeux sur les libertés humaines, les droits élémentaires comme le droit à un traitement équitable et le droit à la santé.

Quels sont les biais les plus difficiles à traiter dans les algorithmes ?

Les biais les plus difficiles à traiter sont ceux dont on ne s'aperçoit pas. Ceux que l'on ne veut pas voir. Nous sommes tous des « algorithmes » humains biaisés : on réagit avec notre subjectivité, notre vision du monde et l'on fait des choix subjectifs dépendant de notre culture, de notre éducation, de note histoire personnelle. L'humain est par nature biaisé. Comme je vous le disais, à mon sens, l'IA ne crée pas de nouveaux biais, mais elle met en valeur la diversité des biais existants. L'enjeux de l'ingénieur, du chercheur, du programmeur ou d'une instance de régulation, est de s'assurer que il y ait une grille de lecture suffisante pour déceler les biais. Une fois que le biais est décelé, c'est comme un conflit d'intérêt : s'il est découvert, on peut le traiter.

Quand le biais est identifié, tout en prenant en compte le contexte et le référentiel culturel, le traiter sera relativement simple. L'enjeu est vraiment de les identifier, de les voir et de ne pas les laisser passer pour des intérêts politiques, économiques, hégémoniques ou autre. Cela pose la question ouverte du besoin d'une réglementation, d'une certification ou d'un standard international pour pouvoir s'assurer qu'à chaque niveau de développement d'une IA, on ait des protocoles pour identifier les biais en pouvant y apporter une résolution. Ce ne sont pas tant les biais que l'on crée qui sont difficiles à traiter, mais plutôt les biais que l'on ne voit pas, car a priori on est capable de les traiter.

Vous avez évoqué les standards internationaux et les protocoles préétablis, identifiez-vous d'autres approches pertinentes pour résoudre les biais dans les algorithmes ?

Il pourrait y avoir des mesures coercitives telle qu'une réglementation, une autorité tierce au niveau d'une région, par exemple à l'échelle de l'Europe ou des USA, qui obligeraient tout développeur ou entreprise construisant des systèmes d'IA à suivre certaines règles de transparence, d'explicabilité, etc.

Ce qui est aussi très efficace, c'est la démarche sur la base du volontariat : les entreprises elles même décident de s'obliger à respecter des standards émis par des comités pluridisciplinaires. C'est ce que fait IEEE, dans la définition de standards qui sont en train d'être mis en place, impliquant des critères éthiques et d'autres que je vous ai évoqués.

Il y a donc une démarche coercitive avec la certification et la réglementation, mais aussi une démarche sur le volontariat au niveau international.

Au sein de votre expérience professionnelle, avez-vous un exemple d'une mesure que vous avez mise en place pour résoudre un biais algorithmique ?

Oui, de manière concrète, dans les sociétés dans lesquelles j'ai pu investir, en m'assurant que les protocoles d'identification et de résolution des biais que je vous ai décrits étaient bien mis en place. Comme j'investis dans des sociétés dans le domaine de la sécurité, du médical et du transport, acquérant des informations sur l'environnement public, il y a bien évidemment des gardes fous à mettre en place sur les enjeux des biais. Je n'ai pas moi-même mis en place et codé ces algorithmes, car j'étais investisseur et membre du board de ces sociétés. En revanche, il y avait un regard très poussé au niveau du management pour s'assurer que les ingénieurs mettaient en place des gardes fous pour vérifier que les données collectées étaient soit anonymisées, soit traitées de manière non biaisée. On vérifiait aussi que les algorithmes déployés étaient fait en faveur de la diversité, de la dignité humaine, etc.

Plus récemment, au niveau de la recherche entreprise au sein de l'AI Institute d'emlyon, le sujet des biais est très important pour voir comment créer de la valeur business en respectant l'humain. Ce que m'ont appris mes 20 ans de carrière dans le domaine de l'IA, de la robotique et des technologies, c'est que toutes ces technologies sont une source de valeurs économiques mais aussi une véritable opportunité au centre des enjeux. Il faut s'assurer et veiller en permanence à ce que ces technologies se développent pour le bien être de l'Homme. Il faut que non seulement les enjeux éthiques, mais aussi les enjeux sociétaux et de valeurs humaines, soient constamment bien mis au centre. Il faut se demander « Est-ce que je suis en train de développer un système qui in fine aide l'Humanité ? », même au niveau de la recherche, du marketing, de la finance, de la supply chain. Cela peut paraître un peu tiré par les cheveux de se demander « Comment la finance doit considérer l'Humain au niveau d'algorithmes de trading ? » mais c'est justement une très bonne question. Par exemple, sur les questions de fraudes, d'identification des flux monétaires ou de prise de décisions sur les marchés ou d'arbitrages. Donc, « Comment placer l'Humain ? » C'est en effet une question qui, soit en tant qu'investisseur, soit en tant que chercheur ou en tant qu'expert-conseil auprès des différentes institutions internationales, j'essaye de mettre au centre de mon action.

Nous arrivons à la deuxième partie des questions, davantage centrée sur l'éthique dans les algorithmes. Comment définissez-vous l'éthique ?

Je ne suis pas éthicien, il faudrait peut-être plutôt poser cette question à quelqu'un qui a vraiment le recul nécessaire. Pour moi l'éthique est très lié aux valeurs humaines. C 'est l'enjeu de l'Être avec un grand « E ». À mon sens, parler de l'éthique d'une technologie, c'est plutôt parler de l'éthique de l'impact de cette technologie sur l'Humain. Ensuite, la question de l'éthique est un vrai sujet pour moi, mais c'est très culturel. Sans parler du Bien et du Mal, on ne va pas traiter de la même manière l'éthique en Asie, en Inde, aux USA ou en Europe, pour des questions culturelles, religieuses ou historiques. Encore une fois, et sans être éthicien, l'éthique est tout ce

qui a attrait aux valeurs humaines et au respect de ces valeurs dans un contexte culturel, historique et religieux donné.

Vous évoquez que l'éthique peut être plurielle et subjective, selon vous, peut-elle être objectivement appliquée dans les algorithmes ?

C'est la grande question que l'on se pose au niveau de l'IEEE, on essaye de voir comment définir et développer un protocole d'algorithme éthique, et ça se passe à différents niveaux. À mon sens, cela passe premièrement par la formation des ingénieurs à ce qu'est l'éthique. C'est primordial, car c'est déjà sensibiliser les acteurs du premier niveau aux enjeux éthiques et ce n'était pas forcément fait avant, c'est assez nouveau. Donc, il faut déjà former ces personnes à la question « C'est quoi l'éthique ? ». Après, il faut identifier les dénominateurs communs que l'on peut retrouver entre les différentes cultures, notamment sur la dignité humaine, la vie privée, etc.

Il y a aujourd'hui une différence entre ce que fait le chercheur et ce qu'implémente le politique. Il y a des chercheurs chinois extrêmement brillants, ayant des vraies notions sur la *privacy* et sur la dignité humaine, qui développent des technologies de reconnaissance d'image ou autre, pouvant être utilisées à des fins politiques détournées. Ce ne sont pas les chercheurs eux-même qui font ça, mais c'est l'utilisation qui peut en être faite par un grand donneur d'ordre. Il faut faire la distinction entre l'utilisation politique d'un système, par exemple en Chine, et ce pourquoi il a été développé et de quelle manière. Il faut donc vraiment trouver les différents acteurs : développeur, intégrateur, utilisateurs finaux, etc. participant à cette chaîne de valeur et identifier suffisamment d'intérêts communs pour dire ce que doivent faire ou ne pas faire ces systèmes. Cela permettra de mettre en place des chartes, des réglementations ou des certifications qui assurent que ces valeurs, que l'on considère universelles, soient respectées. C'est fait par exemple au niveau de la Commission Européenne, qui regroupe quand même 27 états. Même si c'est l'Occident, il y a des différences entre les espagnols, les italiens, les finlandais, les danois, etc. Comment trouver des valeurs communes et définir des critères objectifs pour ces valeurs universelles ? Je pense que l'on peut y arriver si tous les acteurs autour de la table sont relativement responsables, même s'ils viennent de cultures et d'environnement différents. C'est vraiment un très gros enjeu qui n'est pas encore résolu et sur lequel on travaille beaucoup.

Avez-vous identifié d'autres difficultés liées à l'application de l'éthique dans les algorithmes ?

Je ne pense pas qu'il y ait de problèmes techniques à le faire. Votre question est complexe, je n'ai pas de réponse finale. Cependant, il y a déjà des groupes de développeurs qui ont essayé résoudre ces problèmes avec les SALA (Système d'Armes Létales Autonomes), ce que certains appellent les « robots tueurs ». Il y a eu tout un débat depuis 2011 qui est encore courant, entre les acteurs de la Défense, c'est-à-dire les armées, les chercheurs en IA et le public, représenté par des associations comme Human Rights Watch, qui abordent la question « Peut-on coder dans les robots militaires un algorithme éthique ? ». C'est faisable d'une certaine manière, mais est que cela sera suffisant ? Est-ce que l'on pourra prédire tous les cas possibles ? Il est difficile d'y répondre.

Toujours dans le domaine de la Défense, il y a aujourd'hui un code de conduite qui s'appelle le DIH (Droit International Humanitaire). Il régit les conflits entre les différentes nations. Bien évidemment, je ne vais pas parler ici des organismes et des états terroristes qui ne respectent pas le DIH, mais plutôt des états démocratiques, qui représentent la majorité dans le monde, et qui le respectent. Le DIH parle de certaines notions tel que la proportionnalité, c'est-à-dire la

dimension de l'attaque par rapport à la dimension de la réponse, de la capacité de discernement, par exemple entre une cible civile et une cible militaire, etc. Il y a donc des notions assez claires émises par le DIH qui peuvent être intégrées dans les algorithmes. En revanche, il peut arriver qu'il y ait des circonstances de prises de décisions sur le terrain qui sont nouvelles ou complexes, et l'algorithme ne saura alors pas y répondre car ne l'a jamais vu, ni appris, ni codé. Le choix fait par les militaires, qui me paraît pertinent à mon sens, est que l'on garde l'Homme dans la boucle. Cela signifie que in fine, il n'existe pas de système totalement autonome : il y a toujours une supervision de l'Homme. Ce que disent les militaires, c'est : « Man-in-the-loop » et « Man-on the-loop » ou bien « Man-out-of-the-loop ». Cela veut dire que la décision létale, d'action finale, doit toujours être prise par un humain : ces systèmes ne sont pas totalement autonomes, et il n'en existe aucun aujourd'hui qui l'est dans sa prise de décision et dans son action. Le fait d'avoir un humain dans certaines décisions finales critiques permet de s'assurer que tout ce qu'on n'a pas pu coder dans les algorithmes est comblé par une prise de décision dans une chaîne hiérarchique. Ça c'est très important et ça rejoint ce que je vous disait auparavant : les systèmes doivent être faits pour l'intérêt de l'Humanité. Ça peut paraître étrange de parler d'intérêt de l'humanité dans un contexte de défense, mais c'est quand même important et c'est écrit dans le DIH.

Selon vous, qui doit décider de l'éthique à mettre en place dans les systèmes algorithmiques ?

Je pense que ça peut être un comité indépendant et multidisciplinaire. Des chercheurs, universitaires, entreprises, ingénieurs, individus de la société civile, comme Human Rights Watch par exemple, car c'est important que ces associations soient présentes. Il y a aussi les associations de préservation du climat pour s'assurer des impacts sociétaux.

En résumé, il faut une autorité indépendante, pluridisciplinaire, représentative de la société civile dans sa diversité qui pourra vraiment s'assurer de la définition et de la mise en place de ces normes. À mon sens, cela ne peut se passer que par le débat et par la recherche de consensus, ce qui explique que ça soit long à mettre en place et qu'il faut d'abord trouver une définition commune de l'éthique, afin de réfléchir ensemble à comment appliquer les différentes variables.

Y a-t-il des domaines où l'éthique n'est pas importante dans les algorithmes ?

Très simplement, si je qualifie l'éthique comme étant le respect des valeurs humaines et du bien être de l'Homme, c'est bien évidemment important partout. Encore une fois, que ça soit dans la finance, le marketing, la défense, le transport, etc. : à partir du moment où l'on parle du respect de l'Humain, de mon point de vue, il n'y a pas de situation où ce n'est pas nécessaire. Donc oui, l'éthique doit intervenir dans tous les secteurs et à tous les niveaux.

Nous arrivons à la question de conclusion. Quels sont les grands enjeux actuels ou à venir de la recherche concernant l'éthique des algorithmes ?

Les enjeux, ce sont tous les biais dont nous avons parlé avant. Que doit-on mettre en place comme procédure de vérification pour s'assurer que les biais soient identifiés et traités ? Comment le sont-ils ? Soit directement au niveau des algorithmes, soit par la supervision humaine dans des cas critiques comme la Défense.

En outre, il y a la question liée aux valeurs humaines de dignité. Il y a par exemple le cas des robots d'assistance à la personne : dans le contexte actuel tragique des EHPAD, est-ce que la

robotique peut aider les personnes âgées, dépendantes, malades et coupées du lien social, oui ou non ? Les deux réponses sont possibles. Tout ce qui est humain entre dans ces enjeux.

La notion de citoyenneté en fait aussi partie, notamment au niveau de la surveillance. Peut-on développer des systèmes de surveillance généralisés et sans régulation dans un pays démocratique ? Tous ces enjeux sont importants et soulèvent des grandes questions éthiques.

C'est aussi le cas de la question de la responsabilité, lorsqu'un système d'IA commet une erreur : qui est responsable ? Il y donc aussi la question d'identifier la chaîne de responsabilité. Le facteur légal vient après l'éthique mais généralement il en découle.

Les grands enjeux concernant l'éthique sont nombreux, qu'ils concernent les biais, la solidarité, la citoyenneté, la dignité, la responsabilité, etc. Tous ces éléments sont importants et touchent beaucoup de secteurs.

#4. DR. DE LA ROCHE

Date	Length	Communication
May 18, 2020	45’ 10’’	Zoom - Audio call

Pouvez-vous présenter votre parcours ?

Je suis ingénieur de formation, et d’ailleurs, j’ai fait un MBA à emlyon l’année dernière. Avant, j’étais ingénieur en R&D dans le secteur des télécoms. Depuis quatre ans, je suis chez Renault à Sophia Antipolis, près de Nice. La division dans laquelle je travaille se nomme Renault Software Labs : c’est l’entité de Renault responsable des logiciels embarqués dans les véhicules, et elle est divisée en deux parties. La première est située à Toulouse et se charge de la connectivité, c’est-à-dire des connexions 5G et 4G, Bluetooth, interactions avec le smartphone, etc. La deuxième, où je travaille, est la division de Sophia Antipolis près de Nice. Elle s’occupe des logiciels d’aide à la conduite : ce sont toutes les options d’assistance qui vont sortir d’ici quelques années. On est donc impliqué, via la R&D, dans les technologies liées aux véhicules autonomes. Mon poste est celui d’Expert validation et mon équipe s’occupe donc de valider que les logiciels d’aide à la conduite se comportent conformément à nos attentes.

Cette première partie est centrée sur les biais algorithmiques. Comment définissez-vous un algorithme biaisé ?

Le biais est l’un des problèmes assez important qu’il peut y avoir avec tous les algorithmes d’IA, et souvent avec le machine learning. Un algorithme va prendre une décision, et l’on va s’apercevoir que cette décision n’a pas été prise d’une bonne manière car l’apprentissage que l’on a fait de notre solution d’IA n’est pas correcte. Aussi, il faut noté qu’il y a deux sortes de machine learning : le *supervised* et l’*unsupervised*. Dans le domaine automobile, on utilise principalement le machine learning *supervised*, car pour l’instant, l’apprentissage *unsupervised* est trop risqué. En effet, dans le cas de l’*unsupervised*, le logiciel apprend seul, sans que l’humain lui fournisse beaucoup de données d’entrée. Dans le *supervised*, l’humain donne des données pour entraîner l’algorithme afin qu’il devienne capable, à terme, de prendre des décisions basées sur ses connaissances du passé.

Dans le cas du véhicule autonome, si l’on souhaite tester une fonctionnalité de freinage d’urgence, il faudra s’assurer que l’algorithme ne soit pas entraîné seulement à freiner à la vue d’un humain, sinon, lorsqu’il verra autre chose qu’un humain, il ne freinera pas. Le biais vient donc du fait que les données d’entrée n’ont pas suffisamment couverts tout ce que l’on souhaite couvrir, la décision de sortie va donc être mauvaise.

En conséquence, beaucoup de problèmes éthiques sont liés au biais. Par exemple, le grand évènement annuel de la ville de Nice est le carnaval, mais, suite aux problèmes d’attentats, l’année dernière a été mis en place à l’entrée un système de reconnaissance faciale. Cela fonctionnait sur la base du volontariat car il n’est pas encore légale d’utiliser cette technologie sans accord. Tous les participants qui entraient dans la zone du carnaval de Nice se faisaient photographier et étaient *trackés* pendant leurs déplacements. C’est une société israélienne qui produit ce logiciel, et il s’est avéré très efficace. Cependant, ils se sont rendu compte que les personnes à la peau noire étaient mal identifiées, et ça pour une raison très simple : l’algorithme a été principalement entraîné sur des personnes à la peau blanche. C’est un exemple de biais :

l'algorithme ne fonctionnera pas ou mal sur ces personnes car il n'a pas été bien entraîné. Le biais est donc problématique dans la mesure où il peut engendrer des problèmes éthiques.

Quelles sont les sources de biais dans les algorithmes ? Vous avez évoqué les données, en avez-vous identifié d'autres ?

Les biais proviennent principalement des données, mais dépendent aussi de la qualité de l'algorithme utilisé. On peut avoir des bonnes données, mais notre méthode d'IA peut être implémentée d'une mauvaise manière, et ne pas correctement intégrer les données d'entrée, créant des biais en sortie.

Quels sont les biais les plus difficiles à traiter dans les algorithmes ?

Tous les biais sont difficiles à traiter, car souvent, ils sont découverts a posteriori. Dans le cas du machine learning *supervised*, il faut essayer de couvrir tous les cas possibles avec les données d'entrée, mais on peut oublier une situation, quelque chose peut nous échapper, et donc entrainer la création de biais dans notre algorithme. Je ne pense pas qu'il y ait un problème plus compliqué qu'un autre ; ils le sont tous. Il faut essayer en amont d'identifier tous les problèmes possibles afin d'avoir des données d'entrée qui les couvrent tous.

Selon vous, quelles sont les meilleures approches pour résoudre les biais dans les algorithmes ?

Il y a beaucoup de recommandations.

Le premier point, et c'est une certitude, ce sont les données d'entrée : plus elles sont de qualité, exhaustives et diverses, plus on diminue les chances de biais.

Le deuxième point, c'est l'algorithme d'IA au sens strict : il faut bien le valider et le vérifier, pour être sûr qu'il se comporte correctement et n'engendre pas des biais. L'audit et la validation sont des éléments clés ; ils correspondent à l'analyse dans des situations diverses de notre IA pour s'assurer de ses bonnes réactions.

Une autre méthode, n'ayant pas connaissance de toutes les entrées à prévoir au moment de l'apprentissage, consiste à réaliser des sondages et à interroger des gens : il faut impliquer beaucoup d'individus. Celui qui développe la solution ne doit jamais être seul, tous les acteurs sur qui l'algorithme aura un impact doivent être impliqués, et le public notamment. Dans le cas des voitures autonomes, on va par exemple faire des sondages auprès des clients : « *À votre avis, comment doit se comporter le véhicule dans telle situation ?* », « *Quelle est la bonne situation dans tel cas ?* », etc. Cela permet de s'assurer qu'on couvre un maximum de problèmes et de choix possibles.

Du côté de la recherche, il y a aussi beaucoup de R&D qui est faite en IA pour améliorer la qualité des algorithmes. En ce moment, un sujet important pour les chercheurs concerne l'explicabilité des algorithmes, consistant à rendre l'IA explicable le plus possible. Si l'on arrive à expliquer comment les décisions sont prises, on pourra réduire plus facilement les biais en choisissant mieux les données d'entrée.

Un autre point important est celui de la formation : il faut former les gens qui implantent les IA aux problèmes éthiques liés au machine learning et aux biais. Il faut qu'ils soient au courant, afin qu'ils puissent tout mettre en œuvre pour les éviter.

Selon moi, ce sont les principales approches, et surtout, il ne faut ne pas laisser quelqu'un seul implémenter une solution car le risque est trop élevé : il faut travailler collectivement.

Au sein de votre expérience professionnelle, avez-vous un exemple d'une mesure que vous avez mise en place pour résoudre un biais algorithmique ?

Il y a tellement de situations possibles dans l'aide à la conduite qu'on ne peut pas tout valider sur route avec des véhicules réels. D'une part, cela serait trop dangereux car une personne qui prendrait la route avec un algorithme en développement mettrait sa vie et celles des autres en danger. D'autre part, cela prendrait des années pour couvrir tous les cas possibles en condition réelle. De fait, il y a des validations faites sur véhicules réelles, mais une grande partie est faite sur des logiciels de simulations, qui simulent un véhicule et un environnement. Cela permet de créer de nombreuses situations et de vérifier que le comportement est bon. Un exemple de biais qui nous est apparu est celui de la météo : on validait trop souvent dans des situations avec des bonnes conditions météo. Cependant, s'il y a de la pluie par exemple, les capteurs réagissent différemment et les données ne seront pas les mêmes car il est évident que la caméra ne verra pas aussi bien. On a donc mis en place des simulations pour que nos véhicules soient adaptés à des situations de neige, de pluie, etc.

Nous arrivons à la deuxième partie des questions, davantage centrée sur l'éthique dans les algorithmes. Comment définissez-vous l'éthique ?

Lorsque l'on parle d'éthique, on parle d'avoir un bon comportement. Un système éthique est un système ayant un comportement moral, qui ne prend pas de décision qui serait contraire à ce qu'un humain normal aurait voulu faire. Le cas des véhicules autonomes et d'ailleurs très intéressant au niveau de l'éthique car il y a de nombreux cas extrêmes. Par exemple, un véhicule détecte deux obstacles : un chien et un homme, mais l'algorithme sait que même avec un freinage maximal, la distance de freinage va être supérieure avec la distance jusqu'aux objets, et qu'il y aura donc forcément une collision. Il va falloir qu'il choisisse : faut-il rester dans ma voie pour écraser l'un, ou changer de voie et écraser l'autre ? Si l'algorithme décide d'écraser l'humain et de sauver l'animal, on peut considérer que ce n'est pas une décision éthique. Quand on parle d'éthique, on parle d'essayer de faire du mieux possible, de faire ce qu'aurait fait un adulte moral normal. Ce qui est compliqué avec l'éthique, c'est qu'elle n'est pas universelle, elle est différente suivant les cultures, les religions, les pays, etc.

Selon vous, l'éthique peut-elle être objectivement appliquée dans les algorithmes ?

Ce n'est pas facile, mais c'est l'objectif d'arriver à cela. Dans le cas d'un véhicule vendu comme totalement autonome, où toutes les actions sont prises par le logiciel et où l'humain n'a aucun contrôle, le responsable principal devient le concepteur. S'il y a un accident, alors il y aura un procès et cela peut coûter très cher à l'entreprise. On essaye donc de faire en sorte que le logiciel prenne une décision moins pire que ce qu'aurait fait un humain. S'assurer que le logiciel fasse mieux que l'humain, c'est une première piste. On sait qu'à grande échelle, il y a eu beaucoup d'études à ce sujet, le véhicule autonome sauvera des vies : les capteurs rendent les logiciels beaucoup plus rapides et capables d'analyser beaucoup plus d'informations qu'un cerveau humain. En général, on sauvera donc des humains, mais c'est dans les cas critiques et extrêmes que les constructeurs essayent de s'assurer au mieux que la décision prise soit meilleure que celle qu'aurait pu prendre un individu. Pour ça, il faut beaucoup de validations, de données d'entrée, de situations prévues, de R&D, de personnes impliquées, etc. Ce n'est pas un problème simple.

Y a-t-il des domaines où l'éthique n'est pas importante dans les algorithmes ?

Oui, il y a de nombreuses solutions d'IA ou l'éthique n'a pas ou très peu d'importance. Par exemple, il y a plein d'applications sur smartphone pour faire de la reconnaissance de code barre et comparer le prix des produits entre les différents magasins concurrents. S'il y a une erreur, ce n'est pas très grave : l'utilisateur payera peut-être un petit peu plus cher mais il n'y aura pas d'implications humaines ni de vies en jeu. En revanche, certaines applications d'IA impliquent des vies humaines : on a évoqué les voitures autonomes, mais dans l'armement aussi. Par exemple, les missiles à têtes chercheuses ont des vies en jeu.

Selon vous, qui doit décider de l'éthique à mettre en place dans les systèmes algorithmiques ?

Quand une application utilise de l'IA, il faut essayer d'identifier tous les acteurs impliqués. Dans le cas d'une voiture autonome, il y a de nombreux acteurs : le constructeur, les fabricants de pièces, les gens sur la route, le vendeur de la voiture, les compagnies d'assurance, le conducteur, l'État, etc. Tous ces acteurs entrent en jeux. Même si c'est un sujet récent, il y a actuellement beaucoup de discussions et d'initiatives pour rassembler ces gens et ainsi identifier les décisions à prendre dans les différentes situations. Par exemple, au niveau de la Commission Européenne, il y a *AI Alliance* qui réunit des personnes travaillant dans le secteur de l'IA pour discuter ensemble de ce qu'il faudrait faire pour que les algorithmes aient des conséquences positives. Dans de nombreux pays, il y a plusieurs organisations qui se mettent en place, comme le CERNA en France qui permet d'identifier les bonnes solutions. On voit de plus en plus d'initiatives, mais c'est très nouveau et ce sont des situations complexes regroupant des cas très divers et de nombreux acteurs. Il ne faut pas s'attendre à une solution miracle qui fera l'unanimité et qui plaira à tout le monde, mais il faut quand même essayer de coordonner tout le monde au mieux.

Au sein de votre expérience professionnelle, avez-vous un exemple d'un challenge éthique lié aux algorithmes auquel vous avez fait face ?

Oui, il y en a beaucoup. Je vais vous donner exemple. Vous le savez, plus on entraîne un algorithme, plus il va être bon. Pour cela, il faut beaucoup de données, mais ces données doivent être sécurisées. La question se pose alors de ce que l'on doit faire de toutes ces données collectées. Les capteurs absorbent beaucoup d'informations : ce peut être des données personnelles et privées du conducteur, mais pas seulement. Par exemple, une voiture autonome collecte beaucoup de données sur son environnement, et elle pourrait potentiellement voir que quelqu'un l'a doublée à 180 km/h. À l'avenir, la police pourrait vouloir obtenir ces données afin de donner des contraventions. Avec toutes les données stockées dans une voiture on pourrait faire beaucoup de choses. Il y a aussi des menaces de piratage : la cryptographie et la sécurité de ces données sont donc des enjeux éthiques importants.

Nous arrivons aux questions finales d'ouverture. On parle souvent de l'équilibre entre humains et machines à mettre en place dans le processus de décision, quel est votre point de vue ?

Dans l'état actuel de la science, oui, il faut qu'un humain soit dans la boucle. D'ailleurs, c'est le cas pour les voitures autonomes, même pour les véhicules les plus avancés qui sont commercialisés aujourd'hui, comme les modèles de Tesla, le conducteur est toujours responsable. Il doit avoir les mains sur le volant, surveiller et reprendre la main en cas de problème pour éviter les accidents. C'est comme ça que les constructeurs automobiles se protègent, mais la législation française interdit aussi que la personne au volant dorme par

exemple. L'état du machine learning et les capacités actuelles des véhicules ne sont pas suffisantes pour conférer une autonomie totale : aujourd'hui, oui, il faut que l'humain puisse garder la main.

Cependant, on poursuit cet objectif de se diriger vers des IA de plus en plus autonomes. Que ça soit dans dix ans ou dans vingt ans, on imagine que les véhicules seront totalement autonomes, il y a d'ailleurs déjà des prototypes de véhicules sans volant. Quand on sera à ce stade, l'éthique sera seulement dans l'algorithme, l'humain n'aura probablement pas la possibilité de reprendre la main, mais on doit faire encore beaucoup de progrès en éthique et en machine learning. Ça parait être de la science-fiction, mais c'est bien la direction dans laquelle on se dirige.

Quel est votre point de vue sur la place des biais dans les algorithmes pour la société à long terme ?

Pour l'IA et les biais en général, nous l'avons évoqué, beaucoup de gens travaillent actuellement sur ces sujets. Les algorithmes sont encore très imparfaits aujourd'hui, mais je pense qu'à terme on va réussir à résoudre beaucoup de ces problèmes de biais. Au niveau de la société, il faut réfléchir à la place que va prendre l'IA, car dans de nombreuses situations, les humains ne serviront plus à rien. Par exemple, pour les voitures autonomes, il n'y aura plus besoin d'apprendre à conduire ni de permis de conduire. La machine va prendre une place croissante, et l'Homme va se laisser dépasser par certaines situations car celle-ci fera beaucoup de choses à sa place : elle va le remplacer dans de nombreux domaines.

Certains ont peur car ils pensent que les machines vont prendre le dessus sur les humains, mais ce n'est pas mon avis. Les machines sont conçues par des humains, et ils resteront évidemment toujours supérieure à elles. Potentiellement, cela va améliorer la société, la machine va remplacer des choses qui ne sont pas très intéressantes pour les humains, et ainsi lui permettre de trouver sa place pour de nouvelles choses. Au lieu de travailler à la chaîne dans une usine, ils pourront exercer une activité plus intéressante, comme développer des nouvelles machines. Je pense que cela va élever l'humain en le libérant des tâches qui ne sont pas intéressantes pour lui.

Quels sont les grands enjeux actuels ou à venir de la recherche concernant l'éthique des algorithmes ?

C'est lié à tout ce que l'on a évoqué. Au niveau technique, il y a encore beaucoup de R&D à faire pour réduire les biais et améliorer les algorithmes. Ensuite, il faut réunir les acteurs. En outre, il faut gagner la confiance du public car les gens ont peur de l'IA, même si avec le temps elle est de mieux en mieux acceptée, surtout par les jeunes générations. C'est d'ailleurs pour ça qu'il y a toutes ces initiatives en cours pour éviter de créer des situations négatives, et éviter que l'IA soit rejetée par la société. D'ailleurs, les GAFA ont tous un comité d'éthique depuis un ou deux ans ; cela montre que les grands acteurs sont au courant des enjeux potentiels, et qu'ils embauchent des gens pour réfléchir à ces questions pour aller dans la bonne direction.

#5. MRS. LAIR

Date	Length	Communication
May 8, 2020	57’ 00’’	Zoom - Audio call

Pouvez-vous présenter votre parcours ?

J’ai étudié à emlyon et en suis sortie en 2011. Cela fait maintenant neuf ans que je travaille. J’ai toujours travaillé dans les nouvelles technologies et je suis arrivée sur les sujets d’IA il y a trois ans en co-fondant l’association Women in AI. J’ai ensuite rejoint Snips, qui a été racheté récemment par Sonos, qui propose une solution d’assistance vocale qui fonctionne entièrement sans le *cloud* mais en local dans les objets. Cependant, le sujet des biais était beaucoup plus fort chez Women in AI mais se retrouvait aussi chez Snips. Récemment cette année, j’ai créé mon entreprise autour de l’intelligence artificielle responsable et du *AI for good* : elle se nomme The Good AI. Elle a pour vocation d’être une plateforme de ressources pour les gouvernements, les entreprises et le secteur publique afin de les aider et de leur apporter des outils pour construire des modèles qui soient éthiquement viables.

Cette première partie est centrée sur les biais algorithmiques. Comment définissez-vous un algorithme biaisé ?

Un algorithme est un ensemble de règles mathématiques mis en place pour arriver à un résultat qui dépend de plusieurs variables. Biaisé veut à la fois tout dire et ne rien dire, car la société est biaisée et l’humain est biaisé à la base. Un algorithme biaisé est celui qui donne une image de la société à un moment donné, ce n’est pas forcément positif ou négatif, ça retranscrit normalement les tendances que l’on retrouve dans les décisions humaines. Un algorithme, si l’on ne fait rien, est naturellement biaisé car nous sommes biaisés. Le rendre non biaisé, c’est presque contre-culturel, voire contre-nature, car il faut réfléchir, en concevant l’algorithme et en utilisant les datasets, pour que ni l’algorithme en lui-même ni les données qui vont le nourrir ne soient pas biaisés. C’est tout un travail qui est non-culturel. Un algorithme biaisé est simplement un algorithme utilisé car ils le sont tous à l’origine.

Vous avez évoqué les données, identifiez-vous d’autres sources de biais dans les algorithmes ?

Il y a des biais au niveau des bases de données utilisées pour nourrir l’algorithme. Par exemple, les bases de données d’images utilisées en computer vision : généralement ces datasets sont américains donc ils reprennent des images américaines. Ils sont donc biaisés naturellement car si l’on utilise cette base de données en Chine, on aura des résultats qui n’auront pas forcément de sens. Il y a donc ces biais de bases de données, qui sont statistiques.

Ensuite, il y a également des biais cognitifs et des biais économiques. Les premiers sont inconscients et se retrouvent chez les programmeurs, lorsqu’ils développent des algorithmes en projetant leur propre vision de la société. On retrouve souvent dans la presse cette image du programmeur blanc, américain et d’une trentaine d’années qui va projeter sur son algorithme sa propre vision du monde. Les biais économiques se produisent lorsque l’on utilise un algorithme avec un objectif business, lorsqu’ils sont programmés pour faire un compromis entre coût et bénéfice. Par exemple, un algorithme achetant des espaces publicitaires pour diffuser des offres d’emploi en ligne, va se rendre compte qu’acheter un espace pour diffuser auprès d’hommes jeunes est moins cher que diffuser auprès de femmes moins jeunes. En conséquence,

l'algorithme va décider de diffuser en priorité auprès d'hommes jeunes : c'est un biais économique.

Pour résumer on a donc des biais statistiques, cognitifs et économiques. Je précise qu'il existe de nombreux biais statistiques possibles, partagés entre les biais de données et les biais de l'algorithme.

Quels sont les risques liés à une absence de traitement des biais dans les algorithmes ?

Les algorithmes existaient avant que l'on parle d'IA et ils ont toujours eu des biais. Ce qui est dérangeant avec l'IA aujourd'hui, au-delà de l'algorithme, c'est qu'elle est automatisée lorsqu'elle est déployée sur le marché. Les *black boxes* sont aussi dangereuses, car, si l'on ne sait pas comment l'algorithme a pris sa décision, alors on peut automatiser des biais de société. Ces biais, on aurait pu les remettre en cause s'ils étaient arrivés sans IA, mais là, ils sont automatisés et amplifiés. Au lieu d'être une simple répétition d'un comportement discriminant, ils deviennent automatiques et accentués : ils mènent vers encore plus de discrimination.

À ce propos, je vous recommande de consulter le *Trustworthy AI* de la Commission Européenne, pionnier en termes de recommandation sur l'IA émanant d'institutions publics en Europe. Ce document reflète la culture européenne dans ce secteur, défendant notamment les *explainable AI*, c'est-à-dire des algorithmes transparents que l'on est capable d'expliquer en évitant les *black boxes*. Lorsqu'une entreprise vend un algorithme à un client, il faut qu'il puisse comprendre ce qu'il y a dedans et comment il a été fait, afin de pouvoir le corriger si les décisions qu'il prend ne sont pas bonnes. Par exemple, la protection sociale américaine utilise des algorithmes pour déterminer les droits aux prestations sociales. Il y a eu ainsi plusieurs scandales d'aides coupées à causes de biais discriminants. Cela montre l'importance de pouvoir comprendre et expliquer les algorithmes sur le marché afin de lutter contre les biais, au risque d'amplifier et d'automatiser des discriminations.

Vous-avez évoqué les *explainable AI*, identifiez-vous d'autres approches pertinentes pour résoudre les biais dans les algorithmes ?

Il y a, selon moi, des approches techniques et des approches non-techniques.

Concernant les approches non-techniques, il faut d'abord éduquer la population à ce qu'est un algorithme, une IA, et comment ça fonctionne. Il faut familiariser les gens avec ces notions, afin qu'ils en soient conscients et aient les bonnes connaissances. Par exemple, un département marketing qui commande un algorithme doit être capable de comprendre ce sujet pour lire l'algorithme et vérifier qu'il n'y a pas de discrimination ou de biais. Il faut donc d'abord éduquer.

Ensuite, il faut introduire de la diversité dans les équipes qui conçoivent les algorithmes. Aujourd'hui, ce sont souvent les mêmes profils, et notamment des hommes. C'est un milieu qui souffre beaucoup du manque de femmes. Ces hommes reproduisent donc une certaine vision dans leurs algorithmes, au-delà du genre. Souvent ils sont jeunes, caucasiens, et reproduisent leur vision incomplète du monde. Plus on aura des équipes diversifiées avec des femmes et des cultures différentes, moins on aura de risque, ou en tout cas on aura une prise de conscience de devoir représenter une diversité. Si c'est un même groupe de personnes qui font les algorithmes, alors ils ne vont pas se rendre compte que ce qu'ils font n'est pas normal. Par exemple, il y a eu une polémique autour de l'application de santé « Health » d'Apple, qui n'a été faite que par des hommes ; ils ont oubliés que les femmes avaient leurs règles. S'il y avait eu une

seule femme dans l'équipe, ça ne serait pas arrivé. La diversité dans les équipes est très importante.

Toujours dans les mesures non-techniques, les entreprises pourraient créer des postes, soit à plein temps, soit en étant un rôle partiel, afin de vérifier que les algorithmes créés soient éthiquement viables. Il y a de plus en plus de postes dans les *responsible AI*. Par exemple, il y a une équipe chez H&M qui travaille sur ces sujets. Certains grands cabinets de conseil mettent aussi en place des practices de *responsible AI*. Il y a évidemment aussi toutes les aspects légaux, de régulation et de réglementation. Aujourd'hui, il n'y a pas encore grand-chose, car c'est un sujet jeune et complexe sur lequel la recherche a encore beaucoup à faire. De plus, il ne faudrait pas brider l'innovation dans les grands groupes : il n'y a pas de réglementation à juste titre, mais il y a des recommandations émises par des institutions, comme le *Trustworthy AI* que nous avons évoqué, la CNIL en France, etc.

Pour résumer les approches non-techniques, on devrait avoir : une éducation, plus de diversité, des nouveaux postes ou tâches dans les entreprises et de la réglementation.

Concernant les approches techniques, il y a des nombreux outils qui existent. Par exemple, IBM a créé un outil nommé *AI Fairness 360°*, défendant le fait d'avoir une solution pour identifier les biais et dé-biaiser les algorithmes. On est encore au stade de recherche, mais des grands groupes produisent déjà des solutions. On n'en connait pas encore bien la valeur, mais nous sommes aux prémices, ça commence.

Le dernier élément est le financement de la recherche, afin de travailler davantage à dé-biaiser les algorithmes, à avoir des moyens, des outils et des méthodes pour construire des algorithmes neutres respectant l'équité algorithmique, c'est-à-dire vierges de tous biais. Cette dernière notion est très ambitieuse car c'est très complexe et ça nécessitera beaucoup d'études ainsi qu'une dynamique commune de la recherche. L'enjeu est énorme, d'une part car si la recherche n'est pas fortement mobilisée, alors les GAFAM, qui avancent sur ces sujets, auront le monopole sur ces outils. D'autre part, car maintenant que la conscience du problème est répandue, si on arrive à dé-biaiser ces algorithmes, l'opportunité est immense. L'IA pourra changer notre société en l'affranchissant des décisions humaines biaisées : allocations de prêts, recrutement, décisions de justice, etc. On pourra éliminer beaucoup plus vite les biais par cette voie qu'on attendant que les mentalités changent, car, par exemple, le World Economic Forum a estimé qu'il fallait, si l'on conservait le rythme actuel, attendre cent ans pour atteindre la parité des sexes dans les domaines de la politique, de l'économie, de la santé et de l'éducation. Ce n'est pas utopiste, mais il faut se saisir de la question.

Quels sont les biais les plus difficiles à traiter dans les algorithmes ?

Je pense que les clichés sont tenaces, par exemple les biais de genre sont ancrés depuis des siècles. Il y a toujours des écarts de salaire et des discriminations à l'embauche. Plus les biais sont ancrés dans la société, plus ils sont difficiles à éliminer. En revanche, lorsque l'on doit rééquilibrer des datasets, cela me paraît plus facile. Par exemple, l'algorithme de recrutement d'Amazon s'était avéré sexiste car il avait été entraîné sur les CV des employés qui sont majoritairement masculins. Ici pour rééquilibrer, il suffit d'ajuster artificiellement la base de données, ce n'est pas très compliqué. Ce qui est plus difficile, c'est lorsque l'on touche à des éléments propres à une certaine culture. Si par exemple demain on utilise en Occident un réseau social chinois, ça peut être très difficile pour que les algorithmes aient du sens. Par exemple, les chinois acceptent en partie d'être surveillés alors que les américains et les

européens rejettent massivement cela. Les relations à la liberté et à la vie privée peuvent être différentes. C'est difficile de « coder » la culture et de l'appréhender, les biais culturels sont complexes. Pour résumer, quand ce sont des biais statistiques, c'est plutôt facile à corriger, quand ce sont des biais cognitifs ou culturels, c'est plus difficile.

Au sein de votre expérience professionnelle, avez-vous un exemple d'une mesure que vous avez mise en place pour résoudre un biais algorithmique ?

Chez Snips, qui créait une solution d'assistance vocale, on passait par des plateformes de crowdsourcing, comme Amazon Turk, pour entraîner notre algorithme : Amazon paye des gens pour prononcer des mots ou des phrases. Les enregistrements d'humains, qui constituent la base de données, doivent représenter une diversité de voix, masculines et féminines. Par exemple, lorsque l'on faisait des tests, les voix de femmes fonctionnaient moins bien. Les catégories de personnes qui produisent ces enregistrements doivent donner une juste représentation de la société. Par exemple, en travaillant sur une interface allemande, il faut avoir une juste représentation de la société allemande. Il fallait donc que l'on sache ce qui composait notre base de données, quand et comment elle avait été produite. Ici, ce n'est pas forcément technique, c'est simplement être conscient de ce que l'on fait et des données que l'on utilise.

Nous arrivons à la deuxième partie des questions, davantage centrée sur l'éthique dans les algorithmes. Comment définissez-vous l'éthique ?

C'est très difficile à définir. L'éthique, selon moi, est ce qui est considéré comme moral par un individu. C'est quelque chose de très subjectif d'une culture à l'autre : il y a des ères culturelles avec leur éthique, comportant un ensemble de règles considérées comme morales et acceptables qui permettent en partie d'unir la société. C'est un ensemble de consignes et de principes moraux.

La difficulté des biais éthiques en découle, car d'une ère culturelle à l'autre il y a des différences, alors que certains algorithmes sont communs à ces cultures. Par exemple, ce serait dangereux de demander à une entreprise américaine de fournir à la France un algorithme de justice prédictive car nos codes éthiques liés à la justice sont différents. Il faudrait alors que ça soit produit localement. Pourtant, c'est contraire à la dynamique de business car une entreprise n'a pas envie de se limiter à son marché national. Une entreprise française ou américaine a généralement pour ambition d'exporter ses services à l'étranger. Ici, business et éthiques sont dans une dialectique qui n'est pas complètement incompatible, mais qui est compliquée. Une solution pourrait être de remettre de l'humain dans la boucle ; c'est le concept du « *man-in-the-loop* », au lieu d'avoir un algorithme qui décide de tout dans le processus de décision. Même si l'humain apporte ses propres biais, il peut se faire simplement « conseiller » par l'algorithme.

En conséquence, pensez-vous que l'éthique puisse être objectivement appliquée dans les algorithmes ?

S'il n'y a pas de biais, ça me paraîtrait logique que ça soit objectif, mais c'est une question très complexe qui est toujours à l'état de la recherche, car, par ailleurs, l'équité est également un concept culturel. C'est ce problème qui met en évidence la nécessité d'avoir des algorithmes par ère culturelle, au lieu d'avoir une éthique impérialiste s'imposant à tout le monde.

Avez-vous identifié d'autres difficultés liées à l'application de l'éthique dans les algorithmes ?

Indépendamment des biais, il y a la question de la charge environnementale liée aux IA. Ces systèmes sont très polluants et sont aussi, de ce point de vue, éthiquement discutables. Les solutions reposent d'ailleurs en partie sur les IA elles-mêmes pour savoir comment créer des IA plus durables.

Selon vous, qui doit décider de l'éthique à mettre en place dans les systèmes algorithmiques ?

La recherche académique tel que les chercheurs et les philosophes ont la science sur ces sujets, tant au niveau technique que philosophique, car ce sont des sujets très larges. Ils pourraient créer des conseils consultatifs, en faisant émerger, par exemple, des labels validant certaines mesures et certains types d'algorithmes.

Plus concrètement, les grands groupes songent déjà à créer des labels et mettre en place certains principes, mais leur recherche du profit peut aussi générer des conflits d'intérêts.

Ensuite, les institutions publiques et les gouvernements, qui consulteront probablement les chercheurs, vont également réguler et imposer leurs principes.

Il y aurait donc les académiciens, les entreprises et les institutions.

Au sein de votre expérience professionnelle, avez-vous un exemple d'un challenge éthique lié aux algorithmes auquel vous avez fait face ?

Chez Snips, nous avions un parti pris, celui de faire fonctionner notre assistant vocal en local dans les objets, et c'est pour cela que j'avais rejoint cette entreprise. Les assistants de Google ou Amazon envoient les données vocales sur leurs serveurs pour les traiter, ce qui n'est pas optimal concernant le respect de la vie privée. Snips proposait donc cette alternative, par l'analyse locale de la voix, en respectant donc la vie privée des gens : c'est la notion de *privacy by design*. Ce n'est pas exactement un challenge, mais c'est un choix que j'ai fait en fonction de mes convictions : c'est une solution souveraine, française et qui respectait la vie privée.

Y a-t-il des domaines où l'éthique n'est pas importante dans les algorithmes ?

Il y en a, mais je vais vous répondre à l'inverse, là où l'éthique est très importante : prêts bancaires, prestations sociales, soins médicaux, accès au recrutement, éducation, finance, justice, transports, défense, etc. Dans ces domaines, l'éthique doit être présente partout afin d'éviter tout forme de discrimination. Mais par exemple, dans l'aérospatiale, dans la maintenance prédictive ou dans les domaines industrielles et d'ingénierie pure, indépendamment de la question environnementale, a priori je ne vois pas d'enjeux éthiques. Il y a finalement assez peu de domaines ou l'éthique n'est pas importante, mais quand l'humain ou son bien-être directe n'est pas touché, il y peu d'enjeux. D'ailleurs, c'est dans les domaines du recrutement ou de la justice que les algorithmes ont fait polémiques, alors qu'ils sont présents depuis plus longtemps dans l'industrie ou dans la construction.

Nous arrivons aux questions finales d'ouverture. Plus tôt, vous évoquiez le concept de *man-in-the-loop*. Quel est votre point de vue sur l'équilibre entre humains et machines à mettre en place dans le processus de décision ?

Dans les secteurs que je viens d'évoquer, ceux ou l'éthique est critique, alors il faut de l'humain dans le processus. Dans ces situations, il faut plusieurs individus, des comités consultatifs diversifiés, en phase avec la population qui sera touchée par l'algorithme. Quand c'est

automatisé, les conséquences peuvent êtres très graves ; il faut donc être très précautionneux avec ces questions. Par exemple, dans les aéroports, lorsqu'ils ont mis en place les machines pour faire les check-in, il y a eu un moment de transition où des employés étaient à côté des machines. Ça peut paraître inutile car c'est assez simple à utiliser, mais cela a aidé les gens d'avoir un humain à côté de la machine pour leur en expliquer le fonctionnement. Cela a permis de faire rentrer ce nouveau procédé dans les esprits des gens, et aujourd'hui il n'y a plus besoin de cet humain. Cette transition avec des Hommes est nécessaire pour vérifier qu'on ne soit pas contreproductif et qu'on ne génère pas des conséquences involontaires. En outre, la notion de confiance est capitale, et c'est bien pour ça que le rapport de la Commission Européenne s'appelle *Trustworthy AI,* c'est-à-dire *IA de confiance*. Pour que les gens acceptent que des décisions soient prises par des machines, il faut de la confiance, et cela passe par la présence d'humains dans la boucle. De plus, cela permet de faire des itérations pour ajuster les dysfonctionnements.

Quel est votre point de vue sur la place des biais dans les algorithmes pour la société à long terme ?

J'ai fait une conférence *TedX* sur cette question, et je suis assez enthousiaste à ce sujet. C'est pour cela que j'ai créé l'entreprise The Good AI. Je pense que nous, les humains, sommes imparfaits, biaisés et limités. Cependant, nous avons à notre disposition un outil extrêmement puissant capable de calculer des quantités énormes de données et de nous montrer la réalité là où un humain ne le peut pas.

Je pense que si nous sommes suffisamment éduqués et conscients des risques potentiels, mais également convaincus des puissantes opportunités de l'IA, on saura l'utiliser à bon escient, pour le bien commun. Aujourd'hui, il y a de nombreux usages de l'*AI for good* : environnement, santé, accès à l'éducation, etc. Il y a tous les jours dans la presse des exemples formidables de ce que l'IA a pu amener en quelques années et qui nous aurait pris des siècle sans elle. De plus, si on arrive à avancer sur l'enjeu de l'équité algorithmique, alors, pour la première fois depuis l'Histoire de l'humanité, on pourrait arriver à prendre des décisions sans discriminer d'autres personnes. L'Homme discrimine encore beaucoup ses semblables : le racisme peut être ancré depuis des siècles et être inconscient. Avec cette équité algorithmique, on pourra régler ces conflits et amener davantage de justice sociale. J'ai d'ailleurs créé mon entreprise en ce sens, pour amener les gens à comprendre ces questions et à utiliser ces technologies pour le bien, car tous les secteurs mettent actuellement beaucoup de ressources dans le développement de leurs IA. Aujourd'hui, on parle même de *Technological Social Responsibility* (TSR), expression popularisée par la cabinet de conseil Mc Kinsey, en plus des initiatives de *Corporate Social Responsibility* (CSR).

#6. DR. MASURE

Date	Length	Communication
May 8, 2020	57′ 15″	Zoom - Video call

Pouvez-vous présenter votre parcours ?

Anthony Masure, je suis responsable de la Recherche à la Haute École d'Art et Design de Genève depuis septembre 2019. Je suis agrégé en Design de l'ENS Cachan et j'ai également une thèse de Doctorat en Design de l'Université Paris 1. Avant d'être à Genève, j'étais Maître de Conférence en Design à l'Université Jean Jaurès de Toulouse. Mes propres recherches s'orientent principalement autour des rapports entre design et code, et sur les enjeux sociopolitiques liés aux technologies numériques.

Cette première partie est centrée sur les biais algorithmiques. Comment définissez-vous un algorithme biaisé ?

Prenons d'abord la définition d'un algorithme et définissons les mots séparément.

Un algorithme est une suite d'actions écrite en vue d'obtenir un résultat. Cependant, la différence avec un programme, et par extension un logiciel, est qu'un algorithme n'est pas rédigé en langage machine. Il a une connotation plus générique et n'est pas directement conçu pour être exécuté par une machine.

Ensuite, un biais peut être compris comme une intention, mais avec une connotation négative : c'est une intention qui va être masquée volontairement ou involontairement dans une procédure. Cela donne, en réunissant les deux, une suite d'actions dont la finalité n'est pas évidente ou peut être différente de celle à laquelle on s'attendrait. Il y a l'idée d'une perte de confiance, d'une intelligibilité partielle, ou d'une intention ambiguë. Attachée à ça, il y a la question de l'intentionnalité ou non dans le biais, ce qui vient complexifier le sujet : certains algorithmes sont écrits directement par des êtres humains, d'autres non, mais une machine a-t-elle une intention ? Aussi, on pourrait réfléchir dans l'autre sens et se demander ce que serait un algorithme non biaisé : un algorithme dépourvu d'intention ? Il y a quand même toujours une intention quelque part, et cette intention n'est pas systématiquement consciente des conséquences. On peut donc se demander si cette notion de biais a vraiment un sens, car il y a nécessairement des biais dans tout ce que créer les êtres humains, et ça dans n'importe quel domaine. Il semble que ça soit plutôt les conséquences des biais qu'il faudrait interroger, plutôt que la notion de biais en elle-même.

Quelles sont les sources de biais dans les algorithmes ?

Je pense que ce sont les marqueurs sociaux. Dans les *gender studies*, on dit souvent que le genre est une construction sociale, comme d'ailleurs la plupart des choses qui nous entourent. Toutes les constructions sociales que l'on a au quotidien dans les différentes sociétés se retrouvent dans les algorithmes, à la différence près que les algorithmes sont écrit par les gens dont c'est la spécialité et le métier. Forcément, les groupes sociaux des programmeurs, informaticiens et ingénieurs qui vont créer ces algorithmes forment globalement une couche sociale plus homogène que la société en général : il y a moins de diversité. La majorité des biais que l'on va trouver dans les algorithmes sont explicables par l'éducation, la culture, etc. Ce

public d'informaticiens, programmeurs, développeurs, est plutôt un public masculin, occidental et aisé, et dont les préoccupations ne recoupent fatalement pas la diversité du corps social. Cela s'accentue lorsque l'on prend l'exemple d'applications qui s'adressent à des millions, voir à des milliards de personnes, où la diversité du public visé est énorme par rapport à celle des concepteurs. Ce décalage entre concepteurs et utilisateurs est problématique : je pense que la situation pourrait être améliorée en partie en augmentant la diversité du côté des concepteurs. Je précise que l'on parle ici des algorithmes écrits par les êtres humains, car ceux écrits par des machines sont un sujet plus complexe, même si à l'origine les procédés permettant de les générer sont eux même écrits par des humains. En conséquence, les biais se retrouvent moins dans l'algorithme et plus dans les jeux de données, eux-mêmes construits en grande partie par des humains.

Quels sont les risques liés à une absence de traitement des biais dans les algorithmes ?

Je pense d'abord à la discrimination et à l'idée qu'un algorithme puisse être qualifié de « raciste ». Cependant, si on dit cela, on prête une intention à un algorithme, à quelque chose d'abstrait ou de mécanique, mais ça ne veut pas dire non plus que les programmeurs sont racistes. Ce sont plutôt des dimensions inconscientes liées à des constructions sociales comme le genre, la classe d'âge, le niveau de vie, etc. Ce n'est pas forcément volontaire, mais l'algorithme peut être ressenti comme tel. Il peut également y avoir des conséquences écologiques, car les machines sont de plus en plus puissantes, mais pas forcément plus rapides comme elles devraient l'être, ce qui peut être relié à un manque d'intérêt pour les questions environnementales.

On remarque aussi des biais économiques lorsqu'il s'agit de duper des personnes pour leur faire acheter quelque chose, je pense notamment aux *dark patterns*. Je ne sais pas si on peut encore parler de biais pour ce dernier élément, car dans ce cas, c'est fait consciemment. Je pense que le mot biais est très lié aux neurosciences, et même si ce sont des recherches intéressantes, j'ai tendance à être critique envers ce domaine qui à une approche très « question-réponse », c'est-à-dire que chaque comportement doit être expliqué par tel neurone ou tel système chimique ou psychique. Cela me paraît assez simpliste, et je pense que l'on gagnerait davantage à intégrer une approche psychologique beaucoup plus large, comme par exemple la psychanalyse. On ne devrait pas se limiter aux neurosciences et à raisonner sous forme causale, l'esprit humain est bien plus complexe. Je ne pense pas que le mot biais soit employé par les psychanalystes, par exemple.

Quels sont les biais les plus difficiles à traiter dans les algorithmes ?

Il me semble que ce sont les biais qui vont à l'encontre du modèle économique du service. Même si on arrive à les expliciter, ces derniers sont compliqués à traiter car il va falloir passer par des chaînes de décisions complexes pouvant remettre en question le modèle de l'organisation. On ne va pas savoir comment les traiter sans affaiblir l'entreprise. C'est une question à propos de laquelle on s'est interrogés avec Hubert Guillaud (rédacteur en chef d'*InternetActu.net*) et Véronique Routin (directrice du développement à la Fondation Internet Nouvelle Génération), au sujet du « rétro-design de l'attention » (2018–2019), où la question des biais se posait évidemment. On avait étudié un panel de services pour voir où se situait concrètement, dans les interfaces, ces accroches ou « encoches attentionnelles ». On s'est ensuite demandés, une fois avoir déconstruit ces éléments, comment reconstruire le service d'une façon plus « éthique ». C'est complexe, car on peut trouver des approches qui

fonctionnent, mais finalement, si l'on veut vraiment que ça marche, on remet en cause le modèle économique de l'entreprise qui est précisément basé sur ce temps de cerveau. On aura beau faire l'interface que l'on veut, si on ne change pas le modèle, alors ça ne peut pas évoluer. Je dirai donc que les biais les plus compliqués à traiter sont ceux qui s'opposent à la culture des services commerciaux. Le levier économique est donc compliqué à lever, mais j'ajoute que le levier politique ou de la croyance le sera également. Par exemple, les gens seront-ils prêts à utiliser des voitures autonomes sachant qu'elles peuvent les tuer, même s'il y a peut-être dix fois moins de chance qu'elles aient un accident qu'avec voiture traditionnelle ? Je ne suis pas sûr.

Selon vous, quelles sont les meilleures approches pour résoudre les biais dans les algorithmes ?

Ces voitures autonomes utilisent du machine learning et nous ne sommes pas capable de comprendre tous leurs raisonnements, et donc de remonter la chaîne de responsabilité en cas d'accident. Qui est responsable alors ? La machine ? Le programmeur ? La société n'est peut-être pas prête à encoder dans ses machines le comportement à adopter en cas d'accident, ni à abandonner l'idée de la responsabilité, même s'il y avait beaucoup moins de morts. La question du « design de la responsabilité » est capitale, et on se rend compte qu'avec les algorithmes auto-apprenants, cette notion de responsabilité se dilue. Cela pose un problème politique majeur car plus personne n'est responsable, et par extension, plus personne ne se sent responsable. Comment remettre de la responsabilité dans les technologies numériques ? Je n'ai pas encore de réponse, mais je pense que c'est une première piste à votre question.

Ensuite, je pense que les biais ne peuvent jamais être vraiment résolus, les neurosciences nous l'enseignent. Même si l'on explique un biais de perception, de compréhension ou d'argumentation, il va toujours être présent. On peut l'atténuer, en discuter, mais c'est illusoire de penser que l'on peut totalement l'effacer, en tout cas à l'échelle individuelle. Il y a peut-être des réponses possibles à l'échelle collective, car nous n'avons pas tous les mêmes biais. La diversité pour laquelle je plaidais au début de l'entretien pourrait participer à contrecarrer en partie ces biais-là.

On pourrait aussi utiliser des mécanismes de vérification comme des check-lists et des processus de qualité. En établissant des typologies de conséquences nuisibles des biais et des publics qui les subissent, cela pourrait peut-être permettre de mettre en place des processus de validation pour les éviter.

Je pense que cela rejoint le fait de rendre les algorithmes « explicables », plus intelligibles, et le logiciel libre à un rôle à jouer là-dedans car on peut inspecter le code source. Cette possibilité devient d'ailleurs impérative dans les institutions publiques, on l'a vu par exemple en France avec la le système d'orientation scolaire APB. Le fait de systématiser la documentation et de la mettre à disposition peut, couplé avec une certaine diversité, contribuer à réduire les biais côté conception.

Au sein de votre expérience professionnelle, avez-vous un exemple d'une mesure que vous avez mise en place pour résoudre un biais algorithmique ?

Avec Hubert Guillaud et Véronique Routin de la Fing, nous avions étudié une application de contrôle parental et mis en évidence plusieurs présupposés dans l'application car elle surveillait à outrance les enfants. Les concepteurs n'avaient pas forcément conscience de toutes les conséquences de ces pratiques, sans doute parce que ça n'avait pas été beaucoup testé sur le

terrain ni comparé avec des théories sur l'éducation. Pour résoudre cela, des étudiants en design de l'Ensci que nous encadrions avaient établi un contre-scénario qui inversait le contrôle parental, c'est-à-dire que les enfants pouvaient aussi exercer un contrôle sur leurs parents pour remettre de la symétrie dans le service. On s'est rendu compte que cette notion de symétrie était assez fertile et pouvait être appliquée à d'autres services. Lorsque l'on demande quelque chose aux utilisateurs, il faudrait lui permettre de demander au service quelque chose en retour. Par exemple, Amazon collecte des données sur les gens, et certes cela peut être un problème, mais il pourrait être moindre si l'utilisateur pouvait y avoir aussi accès pour les intégrer dans d'autres services ou les utiliser lui-même. Le souci actuellement est qu'il y a une asymétrie très grande, car la collecte de données ne sert pas du tout aux usagers. En mettant de la symétrie, on peut réduire en partie certains problèmes et rendre les biais apparents.

Nous arrivons à la deuxième partie des questions, davantage centrée sur l'éthique dans les algorithmes. Comment définissez-vous l'éthique ?

Pour moi c'est un mot problématique que je n'aime pas trop employer, même si c'est un mot vieilli, je préfère le mot de « morale ». L'éthique est quelque chose de vraiment lié à l'être humain, dans une approche personnelle et individuelle alors que la morale a une dimension plus collective. J'ai du mal à imaginer un corps social se basant principalement sur des approches individuelles voire individualistes. Je préfère donc le terme de morale, à condition que nous puissions la discuter et que ça ne soit pas quelque chose de vertical et d'imposé par une organisation.

D'ailleurs, nous n'avons pas encore évoqué les designers mais seulement les concepteurs. Pourtant, le design est important car les algorithmes ne valent rien tant qu'ils ne sont pas intégrés à des services qui ne peuvent pas exister sans interfaces, qu'elles soient vocales, visuelles, ou autres. Des biais peuvent être embarqués dans ces interfaces, et les designers jouent un rôle là-dedans, même si malheureusement, pour le moment ils ne sont pas assez formés au niveau de la psychologie. En UX il y a des cours de psychologie, mais c'est de la psychologie comportementale où l'on crée des personas, des suites d'actions, des scénarios, etc. On réduit l'être humain à une suite de comportements, et cela me paraît dangereux. Justement, l'être humain doit prendre de la distance par rapport aux comportements car c'est un être qui échappe au déterminisme. Réduire l'humain aux comportements, comme le fait le design UX, me semble être une catastrophe, et il n'y a pas assez de critiques de cela. On pourrait résoudre cela en introduisant des cours de psychologies « élargies » dans les formations de design, afin de mieux comprendre les conséquences de leurs actions car les designers sont souvent réduits à un rôle d'exécutants par leur chef de projet. De fait, leur faculté critique n'est pas mise à l'épreuve et parfois ils n'en ont aucune : c'est plus facile de suivre et d'exécuter les ordres. S'ils développaient leur esprit critique, ils seraient plus au fait de ces enjeux. Souvent, ils vont « habiller » un algorithme, mais il ne vont pas s'intéresser à l'architecture des choix ou aux processus de décisions qui conduisent à cela. Alors, ils deviennent complices, volontairement ou non, de tous les biais évoqués auparavant tout en ajoutant les leurs. Par exemple en ce moment il y a le confinement, et dans le e-learning relatif au champ médical, on va retrouver tout le vocabulaire du *serious game*. En créant une application de dépistage d'une maladie, cela ne va pas du tout de soi qu'il faille utiliser des codes visuels et ergonomiques du *serious game* dans la santé. On fait cela car c'est à la mode, car c'est facile et qu'on en maîtrise les codes. Pourtant, si on étudiait la psychologie des patients et qu'on avait une approche « test clinique » couplée à des méthodes de design d'enquête, on arriverait à des résultats bien plus pertinents que des univers esthétiques et ergonomiques avec lesquels on a l'habitude de

travailler. J'ai fait un détour par le design car il me semble qu'on ne peut pas « décorer » comme on veut l'interface de l'algorithme qui lui permet d'agir dans le monde.

Oui, je suis d'accord avec vous. J'ai d'ailleurs intégré à ma revue de littérature une partie sur les biais d'usage, qui sont souvent liés aux interfaces.

J'en profite pour vous donner un exemple intéressant de biais lié aux interfaces. J'ai un collègue à Toulouse, Saul Pandelakis, qui a étudié les applications de gestion des règles. Il a fait une étude de tous les codes visuels de ces applications, et il s'est rendu compte de comment elles propagent de nombreux présupposés liés aux enjeux de genres : couleurs roses, pictogramme, etc. Tout cela participe à créer un système sexiste, ou en tout cas problématique, et les designers n'en ont pas forcément conscience, soit parce qu'ils n'ont pas suffisamment connaissance des enjeux de genres, soit parce qu'on leur a demandé de s'inspirer d'autres applications qui marchaient bien. Le design fonctionne en effet beaucoup par inspiration et par copie, car cela demande plus d'efforts de réinventer les choses. C'est un bel exemple de biais incarné dans les codes visuels.

Quelles sont les difficultés liées à l'application de l'éthique dans les algorithmes ?

La difficulté c'est : qui va décider ? Qui va décider des valeurs qui seront embarquées ? Je pense que le neutre n'existe pas, c'est illusoire et il n'est pas souhaitable de tendre vers cela : c'est impossible d'arriver à des algorithmes neutres ou à des interfaces neutres. La question est donc : quelles valeurs va-t-on mettre dedans ? Qui va décider ? Comment ces valeurs vont être explicitées pour que l'on puisse les critiquer ? L'idée est moins de chercher une absence de valeur, mais plutôt de permettre à ces valeurs d'être exprimées pour pouvoir être réfutées. La question « qui va décider ? » est extrêmement compliquée. On pourrait faire entrer des acteurs de la société civile dans la conception : cela s'appelle le *codesign*, et c'est déjà utilisé en architecture. Par exemple, cela a été utilisé à Châtelet-les-Halles ou la population avait été interrogée. On pourrait s'intéresser à ce genre de démarche pour les appliquer aux outils numériques. Avec la société civile, on pourrait obtenir la diversité qui fait défaut côté conception et ainsi tester le service, ou détecter des conséquences inattendues avant la mise sur le marché. Souvent, c'est cela le problème, on joue à l'apprenti sorcier : même si l'on ne peut pas tout prévoir, parfois on commercialise des programmes à grande échelle où l'on n'a même pas essayé d'anticiper les conséquences. D'ailleurs, il y a un champ du design qui traite de cette question et qui s'appelle le « design fiction » ou « design spéculatif ». Ce domaine ne vise pas à concevoir des nouveaux objets, mais à mesurer les potentialités positives ou négatives, ou à mesurer les imprévus, d'objets techniques avant qu'ils soient proposés au grand public. Cela consiste à élaborer des scénarios en extrapolant les conséquences. Il y a évidemment des questions de moyens, il y a des limites, mais ce design fiction nous permettrait de retravailler ces algorithmes en amont, avant de produire des conséquences délétères.

Ensuite, pour la question des décisionnaires, aujourd'hui ce sont plutôt les programmeurs et les développeurs. J'ai esquissé une piste du côté des designers, mais je pense que se sont souvent ceux qui payent qui décident. Alors, comment faire changer les valeurs des décideurs ? Cela peut se faire par la loi : c'est la réponse de Evgeny Morozov qui est un critique des technologies numériques et qui vise à davantage de régulation côté politique. Il s'approche d'une vision communiste adaptée aux enjeux contemporains. Il y a aussi le fait que les gens font davantage confiance à une entreprise privée qu'à leur gouvernement : on le voit aujourd'hui avec les applications du confinement et la querelle entre l'État français et le consortium Apple et Google

pour le traçage des données. On remarque que les gens font plus confiance à Apple et Google qu'à l'État pour maîtriser les enjeux de vie privée. Cela s'explique notamment par le fait qu'en cas de scandale, y a un enjeu économique énorme pour les entreprises.

Ensuite, il y a la question de savoir si les valeurs peuvent être créées par des algorithmes de machine learning de façon automatique. Il y a un auteur que j'ai beaucoup étudié qui a réfléchi là-dessus, c'est Vilém Flusser. Il est décédé en 1991. Dès la fin des années 1980, il avait déjà travaillé sur comment les machines deviendraient auto-apprenantes, et sur comment la production automatique de programmes par un algorithme ruinait complètement le système de pensée occidentale. Pour lui, le système scientifique traditionnel, les religions, la pensée causale ou la pensée finaliste allaient être totalement remis en question par le développement de la cybernétique. Les programmes allaient créer des nouvelles valeurs et il nous faudrait des nouveaux concepts, des nouvelles façons d'analyser les objets informatiques, car selon lui, les nouveaux systèmes de pensées ne fonctionneront plus du tout. D'après Flusser, on court le risque d'arriver dans une société inintelligible, avec une multiplication d'objets « boîtes noires » qui prennent le pouvoir sans que l'être humain n'en soit conscient. Il prend l'exemple de la photographie argentique, où l'on sait comment ça fonctionne, alors qu'une photographie numérique contient un programme et on ne sait pas du tout quelle est son intention, même si elle ressemble à une photographie traditionnelle. Flusser y voyait donc une tromperie, un mensonge inscrit dans les choses. Pour remédier à cela, il évoquait notamment une solution avec l'art, en faisant faire à ces systèmes rationnels des choses dans une dimension qui échappe à la pensée finaliste de l'algorithme, ou en tout cas à une situation déterminée. Il donnait l'exemple de la photographie car pour lui, l'artiste est celui qui est capable de faire sortir de l'appareil photo des choses qui n'étaient pas prévues. Cela me semble intéressant de voir dans l'art une façon de faire dévier les biais ou les valeurs, une façon d'affronter les algorithmes, ou d'une certaine manière de les critiquer.

Je reviens sur un élément dont vous avez parlé, à propos des décisionnaires de l'éthique dans les algorithmes. Est-ce que vous voyez d'autres acteurs qui pourraient intervenir ?

J'ai notamment évoqué la société civile et les décideurs politiques, mais en raisonnant de façon générale je ne pense pas qu'on arrive à des solutions, il faudrait regarder domaine par domaine. Le problème est que le corps politique a largement abandonné ces questions, que ça soit par manque d'intérêt, de compétences, de connaissances et aussi parce que le domaine du privé à largement colonisé la manière de penser de la démocratie participative. Il faudrait peut-être repolitiser ces questions-là, car tout le monde est en rapport avec des algorithmes chaque jour, pour les impôts, la justice, l'éducation, etc. Ce sont des choses très importantes, ce n'est pas débattu au niveau politique et c'est abandonné à des entreprises privées qui ont d'autres préoccupations que le bien commun. Je rejoins en partie Morozov là-dessus : il faudrait davantage de régulation et de prise de conscience politique. Certains seraient fatalistes et diraient que les GAFA ont trop de pouvoir par rapport aux gouvernements, mais ce n'est pas si sûr. On l'a vu pendant ce confinement, pour le meilleur et pour le pire, en quelques semaines nous avons radicalement changé nos modes de vie. Nous n'aurions jamais cru possible que des personnes politiques puissent prendre des décisions aussi gigantesques pour la moitié de l'humanité, et cela donne d'ailleurs des idées aux écologistes pour prolonger certaines actions. Pour les algorithmes, des décisions importantes pourraient être prises pour faire changer les choses, à condition d'avoir une démocratie plus participative.

Selon vous, l'éthique peut-elle être objectivement appliquée dans les algorithmes ?

Le problème se pose du système de valeurs quand c'est un service à l'échelle mondiale, ou les hiérarchies sont très différentes. On le voit avec les services américains qui sont dans d'autres cultures, cela créer des frictions qui ne sont pas souhaitables. Ce qui serait intéressant, ça serait aussi de documenter le système de valeurs du contexte culturel dans lequel l'algorithme s'inscrit. Un genre de grille, ou d'explicabilité au-delà du champ technique : créer une matrice de valeurs pour de pouvoir situer le contexte dans lequel est diffusé un algorithme. Cela permettrait aux concepteurs de mieux situer leurs actions et de mieux prévoir les conséquences. Je pense qu'ici les psychologues pourraient avoir un rôle à jouer ici.

Y a-t-il des domaines où l'éthique n'est pas importante dans les algorithmes ?

Spontanément je ne vois pas, je dirai non.

Au sein de votre expérience professionnelle, avez-vous un exemple d'un challenge éthique lié aux algorithmes auquel vous avez fait face ?

Dans le design, la question se pose tout le temps, entre faire le bien commun ou attirer le client et chercher la rentabilité. C'est ça d'ailleurs qui rend le design intéressant, c'est son côté « impur ». Le design est né des révolutions industrielles. Il met en évidence, avec la production en série, les problèmes de fabrication des objets, la baisse de la qualité à cause de la perte des anciennes façons de faire. Le design se fait ensuite rattraper et absorber par le capitalisme. Il y a donc toujours cette tension entre les aspects réflexifs, inventifs, voire esthétiques, et les enjeux de rentabilité. Ce champ de tension me passionne en tant que chercheur, donc j'aurais du mal à trouver un projet qui n'a pas ce problème-là, il y a toujours des dilemmes éthiques. Ils ne sont pas toujours exprimés et conscients, mais ce débat alimente constamment les designers. Il m'est presque impossible de trouver un objet ou la question ne s'est pas posée. Un bon exemple est le mobilier anti-SDF où la communauté des designers est très partagée, où l'on se demande comment des gens ont pu les concevoir. Cela rejoint un problème majeur du design, le côté « solutionniste ». C'est pénible de devoir trouver des solutions à tout ; cette volonté du design de vouloir résoudre les problèmes du monde, alors que parfois ce n'est pas souhaitable : il faudrait plus d'humilité et prendre en compte davantage les utilisateurs finaux. Coupler le design avec des approches sociologiques serait fécond et permettrait de mieux affronter ces tensions.

Nous arrivons aux questions finales d'ouverture. On parle souvent de l'équilibre entre humains et machines à mettre en place dans le processus de décision, quel est votre point de vue ?

Je dirais que l'opposition humain-machine n'a pas de sens. L'être humain est toujours entrelacé avec la technique, il façonne des objets qui le façonnent en retour. On ne peut pas séparer l'être humain de la technique qui lui permet de dépasser sa condition initiale. L'histoire entière de l'humanité est une suite d'avancées techniques qui refaçonnent ce que l'on comprend comme étant humain. La machine est ce qui, par contraste, fait exister l'être humain. Si l'on dit que c'est le langage qui caractérise l'être humain, alors quand la machine peut parler on va chercher ce qui dans l'être humain n'est pas uniquement du langage et qu'une machine ne peut pas faire : c'est pareil quand une machine aura des émotions, par exemple. Pour moi, « l'équilibre » n'a pas de sens car il suppose une opposition, et dans ma pensée j'essaye notamment de dépasser cette opposition. Si l'on prend un pacemaker, par exemple, l'être humain est cyborg depuis qu'on peut lui mettre dans le corps des objets de ce type. On le voit aussi avec les assistants

vocaux : ils modifient la façon de parler des êtres humains alors qu'ils sont eux-mêmes créés par des humains. Je ne pense pas que cette question d'équilibre nous aidera à penser des solutions plus souhaitables. Je pense qu'il vaut mieux se demander où sont les problèmes de valeurs, comment les résoudre, quelles sont les conséquences non souhaitées, etc. Ou peut-être faut-il se demander où avoir davantage d'autonomie, et où est-ce qu'il en faut moins ?

Quels sont les grands enjeux actuels ou à venir de la recherche concernant l'éthique des algorithmes ?

Je pense qu'il y a les enjeux écologiques bien sûr. Comment exister en tant que concepteur d'objets, et du coup au lieu de construire des nouvelles choses qui rajoutent de la pollution, comment enlever des objets au lieu d'en créer ? Ensuite, les études de genres sont de plus en plus présentes dans les débats publics. Comment limiter les discriminations liées aux orientations sexuelles ou de genre ? Idem pour la question de la discrimination des personnes racisées. À ce sujet je vous recommande de lire le livre *Justice Digitale* qui traite de la criminalisation automatisée, des cartes et des algorithmes prédictifs.

À titre plus personnel, j'ajouterais aussi les enjeux des algorithmes auto-apprenants, et notamment du deep learning, et leur rapport au design, avec cette multiplication des boîtes noires qui brisent la chaîne de responsabilité car on ne peut pas expliquer précisément comment ils fonctionnent. Faut-il ouvrir ces boîtes noires ? C'est une question complexe. Et plus globalement, il y a la question de la concentration. On parlait de « qui décide ? », on peut aussi se demander : « qui décide pour quel nombre ? ». Dans le cas des services numériques, on peut réfléchir à limiter ou réguler le nombre d'utilisateurs que pourrait avoir une même entreprise. Il y a d'ailleurs des lois antitrust pour cela. On peut se dire, dans la même logique, qu'un service qui dispose de deux ou trois milliards d'utilisateurs est trop concentré, ce qui créé un problème à l'échelle planétaire. On peut aussi réfléchir au nombre de personnes atteintes par les mêmes valeurs qu'un service porte.

Quel est votre point de vue sur la place des biais dans les algorithmes pour la société à long terme ?

Je pense que c'est en partie aux designers et aux concepteurs de faire des choses souhaitables pour le plus grand nombre, et pas toujours pour eux-mêmes. Je pense que la question de l'intentionnalité est centrale, il faudrait réorienter la technique vers des directions soutenables plutôt que de réduire les technologies émergentes de manière anti-technique. Dans le deep learning il y a des directions qui sont positives, même si ce ne sont pas les directions principales. Ce sont aussi aux designers de prendre part à ces directions.

#7. MR. PINEAU

Date	Length	Communication
April 27, 2020	49′ 40″	Zoom - Audio call

Pouvez-vous présenter votre parcours ?

Je m'appelle Karl Pineau et je suis doctorant en Sciences de l'Information et de la Communication à l'Université de Technologie de Compiègne. Je n'ai pas un parcours d'ingénieur et je n'avais jamais étudié à l'UTC avant d'y être doctorant : initialement j'ai fait des études d'Histoire de l'Art et j'ai un profil plutôt littéraire. J'en suis ensuite arrivé à faire de l'Information Communication en Master à l'ENS Lyon. Quand j'étais à l'ENS, on a créé l'association Designers Éthiques pour s'intéresser d'abord à la question de l'impact que pouvait avoir les designers sur les utilisateurs, puis au fur et à mesure, on s'est élargi à la question de l'éthique. À la place du terme « éthique », on peut parler plutôt de responsabilité, si l'on veut utiliser des termes plus pragmatiques, c'est à dire la responsabilité des concepteurs dans les services numériques. Cependant, ma thèse n'a rien à voir avec cela car c'est plutôt une activité annexe pour moi, même si en tant qu'apprenti-chercheur, cela entre dans mon champ professionnel. En revanche, sur la question des algorithmes et des biais algorithmiques, j'ai réalisé mon Mémoire de recherche sur la question de la recommandation par les algorithmes de contenu culturel aux utilisateurs. De fait, j'ai un peu travaillé sur ces sujets même si ce n'est pas au cœur de ma spécialité.

Cette première partie est centrée sur les biais algorithmiques. Comment définissez-vous un algorithme biaisé ?

Il me semble qu'un biais, c'est une déviation du jugement, ou en tout cas un jugement faussé. Cela étant, il peut être faussé de plusieurs manières et peut concerner les humains ou les algorithmes. Personnellement, j'ai surtout travaillé, lorsque l'on parle de biais algorithmiques, sur les biais dans les données, c'est-à-dire le fait que le jeu de données que l'on fournit à un algorithme soit biaisé, notamment en machine learning. C'est le fameux exemple de Google Traduction qui va traduire des mots anglais neutres, au féminin ou au masculin en français, selon leur connotation, par exemple « a nurse », « une infirmière » et « a doctor », « un docteur ». Ça, selon moi, c'est vraiment l'archétype du biais car c'est une déviation du jugement. On n'obtient pas ce que l'on veut retranscrire car il y a un biais dans les données, et ce biais, finalement le jeu de données, est issu de précédents corpus de textes qui ont dû être collectés par Google dans le passé, notamment sur Google Books. Il y a un écart dans la manière dont on pouvait traduire des textes auparavant, et la manière dont on va les traduire aujourd'hui en faisant plus attention à l'inclusion. Du coup, c'est typiquement une déviation du jugement qui n'était pas prise en compte à la base.

Ensuite, on peut parler des déviations du jugement liées aux opérations de programmation en elle-même, c'est-à-dire les opérations que l'algorithme va faire. Ces dernières vont être plus difficiles à évaluer car elles seront davantage liées à la « personne », au sens de celui qui va programmer. Ça peut être soit directement le développeur qui code en tant que tel et qui va faire une erreur ou omettre quelque chose dans le code, soit ça va être lié à des choix plus globaux de conception, qui vont induire des biais dans les données. C'est le cas notamment des algorithmes de prédiction de police, pour lesquels on peut filtrer les crimes et délits en fonction

de certains critères. Évidemment, quand on rajoute des paramètres, on ne va pas obtenir les mêmes prédictions que si on avait utilisé d'autres paramètres.

Quelles sont les sources de biais dans les algorithmes ? Vous avez évoqué les données et la conception, en voyez-vous d'autres ?

Quand on parle de biais algorithmiques, on est effectivement soit sur les données, soit sur les concepteurs. Après, de manière plus générale, il y a également les biais cognitifs : ce sont les biais d'interprétation des humains mais qui se retrouvent moins dans la question des algorithmes. Je vois donc les biais dans les données, les biais de conception au sens des opérations logiques, et également les biais d'interprétation des utilisateurs : ça me paraît être les trois formes principales de biais.

Quels sont les risques liés à une absence de traitement des biais dans les algorithmes ?

On en arrive ici à toutes les problématiques d'éthique. Si on a une vision technopositiviste ou solutionniste, et que l'on considère que la technologie est plus neutre, sous-entendu moins biaisée que les humains, on va produire des résultats qui vont être probablement tout autant biaisés mais dont on va moins se méfier. C'est l'exemple des juges américains dont une étude israélienne avait montré qu'ils étaient plus cléments après l'heure du déjeuner, lorsqu'ils étaient rassasiés, plutôt qu'avant, lorsqu'ils avaient faim et avaient alors tendance à expédier les affaires qu'ils avaient à juger.

Est-ce qu'un algorithme qui remplace les décisions des juges sera moins biaisé ? Probablement pas. Le biais de l'horaire sera éliminé car il est très humain, mais on trouvera d'autres formes de biais dans les données. Par exemple, le fait que les populations afro-américaines seront probablement plus représentées dans les statistiques de personnes condamnées par la justice peut être compris comme un facteur par l'algorithme. Mon directeur de Mémoire, lorsque je travaillais sur les algorithmes de Europeana à la Commission Européenne, m'avait dit qu'il avait fait du clustering pour sa thèse, et que c'était très simple d'en faire. Créer des clusters peut prendre une journée, en revanche, essayer de comprendre la formation des clusters peut prendre trois mois. Je pense que c'est ça la grande problématique des algorithmes, notamment en machine learning ou en deep learning : les algorithmes font des choses que l'on ne comprend pas vraiment. C'est l'exemple de l'algorithme de machine learning à qui l'on a voulu faire reconnaître des éléphants, et qui avait en fait compris que c'étaient des objets sur un fond vert car les photos étaient prises dans la nature, mais il n'avait pas compris la forme de l'animal. Ainsi, si on n'a pas conscience qu'il y a nécessairement des biais, on aura probablement des résultats plus biaisés qu'à la normal. Pour moi, ce n'est pas l'idée de dire que l'on puisse avoir un système non biaisé si l'on fait attention, car je pense qu'il n'y a aucun système technique qui peut ne pas être biaisé, à l'image qu'aucun humain n'est neutre ni objectif. En revanche, on peut essayer, en étant vigilant, de résoudre les biais les plus flagrants.

Selon vous, quelles sont les meilleures approches pour résoudre les biais dans les algorithmes ?

Sur la question des biais, je n'ai pas vraiment réfléchi à la meilleure approche, mais plutôt à des pistes. En premier lieu, je pense qu'il faut adopter un point de vue technocritique, et ne pas penser que la technologie est nécessairement meilleure que l'humain, ou en tout cas toujours capable de produire des meilleurs résultats que l'humain. C'est déjà une base qui permet de prendre du recul et de ne pas faire trop confiance.

La deuxième chose, que je dis souvent aux concepteurs-designers mais qui est aussi valable pour les ingénieurs et techniciens qui vont avoir en charge les algorithmes, est qu'au niveau de l'éthique, on a souvent une vision très conséquentialiste, voir utilitariste. On essaye de pallier au problème le plus rapidement possible, avec le fameux Minimum Valuable Product (MVP). On essaye d'avoir des produits très vites utilisables et très vites rentables car on est dans une logique économique souvent assez forte, alors que l'on devrait avoir une vision plus déontologique : on devrait plus se poser la question de ce que produit notre système. Si l'on regarde ce que fait Designers Éthiques aujourd'hui avec la méthode qui est en train d'être développée sur le design systémique, on peut y retrouver des bonnes pratiques sur les questions de biais, car on essaye de regarder quels peuvent être les conséquences d'un système technique sur autre chose que sur ses utilisateurs initiaux : conséquences environnementales, sociales, politiques, économiques, etc. Cela permet de se poser la question de comment on va identifier les biais, et donc la question de comment on analyse et qualifie les métriques de nos systèmes techniques, comment on juge de leurs performances. Si l'on reprend l'exemple de la justice algorithmique, si la métrique est le temps que l'on met à juger les gens, ce n'est probablement pas la bonne KPI d'un point de vue éthique. En revanche, on peut regarder s'il n'y a pas des biais socioprofessionnels dans les gens qui sont jugés, essayer de regarder où nous amènent les métriques que l'on a choisi.

La deuxième question concerne les conséquences produites par les systèmes qu'on a mis en place. On peut trouver des conséquences indésirables, des externalités négatives, c'est-à-dire des biais qui ne viennent pas du système technique que l'on a développé, soit parce que l'on a des mauvaises données, soit parce que l'on a des biais cognitifs, etc. Je pense que là on a quelques grandes notions assez générales que l'on peut mettre en oeuvre. D'ailleurs, il y a un papier que je relaie régulièrement chez Designer Éthiques : c'est une charte écrite en 1999 au sujet des technologies persuasives et qui donne huit règles déontologiques à suivre. Ce n'est pas tout à fait pareil, car les algorithmes n'ont pas nécessairement pour objectif de persuader, mais on retrouve cette logique de faire attention à ce que l'on crée, et moi c'est surtout ça le message que je porte à Designers Éthiques. Si l'on ne regarde pas ce que l'on est en train de créer, et que l'on a l'impression qu'on est en train de créer quelque chose qui va forcément résoudre des problèmes sans se demander quels autres problèmes ça pourrait engendrer, alors on a inévitablement des problèmes à la fin.

Au sein de votre expérience professionnelle, avez-vous un exemple d'une mesure que vous avez mise en place pour résoudre un biais algorithmique ?

En réalisation d'algorithme pas vraiment. En revanche, dans la réalisation de sondages et d'études quantitatives ou qualitatives, ça arrive souvent. Il est très fréquent, en lançant un sondage, que je me rende compte que c'est biaisé car la question n'a pas été bien formulée. Je me souviens aussi, lorsque j'ai fait mon Mémoire, que j'avais récolté des données quantitatives sur des utilisateurs, et une énorme proportion de données était produite par une très faible proportion de personnes : ce sont les fameuses courses 80/20 de *long tail*. De fait, s'il y a très peu de personnes qui produisent la majorité des données à analyser, alors c'est biaisé. Dans le cadre de la démarche scientifique, c'est assez compliqué de redresser ça car il faut faire plus d'analyses. Ce sont les cas auxquels j'ai pu être confronté.

Nous arrivons à la deuxième partie des questions, davantage centrée sur l'éthique dans les algorithmes. Comment définissez-vous l'éthique ?

Ça c'est la grande question de l'éthique. Pour moi il y a deux compréhensions majeures de l'éthique. D'abord, on peut la comprendre au sens de l'éthique philosophique développée par les philosophes depuis la Grèce Antique avec Socrate, Bentham ou Kant notamment. Les premières formes sont l'éthique des vertus, l'éthique déontologique et l'éthique conséquentialiste. Là, on considère que l'éthique est le chemin que l'on adopte individuellement, mais en se référant à l'une des trois grandes formes pour trouver sa voie idéale dans les actions que l'on doit mener au quotidien vis-à-vis de sa morale. Ça c'est l'approche éthique plutôt philosophique.

Cependant, lorsque l'on parle d'éthique dans le langage courant, je pense que c'est juste « faire le bien ». Lorsque l'on parle d'une entreprise qui est éthique, c'est une entreprise qui fait du bien, ou qui donne le sentiment de ne pas être agressive d'un point de vu marketing ou commercial par exemple. Parfois, je pense que ça tend même à dire que pour être éthique, il faudrait être hors marché économique. Je pense qu'il y a cette distinction de l'éthique d'un point de vue philosophique, et de l'éthique dans le langage courant.

Quelles sont les difficultés liées à l'application de l'éthique dans les algorithmes ?

Je suis contre la formulation assez classique de dire qu'il faut appliquer de l'éthique, car, de mon point de vue, tout le monde a une éthique. Enfaite, quoi que l'on fasse, on agit d'une manière éthique, ce n'est pas forcément conscient mais ça en est une quand même. Selon nos choix, on est plutôt conséquentialiste, déontologique ou éthique des vertus. Je n'ai jamais rencontré de personnes dont on puisse dire absolument ou fermement « cette personne n'a aucune éthique » car en réalité, l'éthique est justement une manière d'analyser une réaction à une situation. Donc, quand on dit qu'il faut qu'un système ait plus d'éthique, ce que l'on veut dire en réalité, c'est qu'il faut qu'il soit plus déontologique, au sens où il faut qu'il fasse plus attention à ce qu'il fait dans ses actions.

En conséquence, pensez-vous que l'éthique puisse être objectivement appliquée dans les algorithmes ?

Je pense justement qu'il ne faut pas dire que le système est « non éthique », mais qu'il s'agit plutôt de dire que pour le bien commun, pour la société au niveau des individus, on reproche au système technique de ne pas faire les bon choix.

Dans une société, on repose sur un modèle de valeurs morales que l'on partage, ou en tout cas sur lequel on est globalement d'accord. Ensuite, dans les actions qui sont effectuées par le système qui est critiqué, celles-ci peuvent êtres catégorisées dans un système d'éthique. Le plus souvent, à mon avis, elles sont classées dans le conséquentialisme, cependant, ce que l'on voudrait vraiment, c'est que ces actions soient plus déontologiques. Par exemple, à mon avis Facebook à une vision éthique conséquentialiste ou éthique des vertus dans la majeure partie de ses actions. Cependant, ce que la plupart des gens critiquant Facebook veulent dire lorsqu'ils disent qu'il n'est pas éthique, c'est qu'ils attendent de Facebook qu'il ne fasse pas certaines actions qu'il peut quand même faire. Ça, c'est précisément de la déontologie, c'est-à-dire s'interdire à soi-même de faire des choses par ce que l'on a décidé préalablement qu'elles étaient mauvaises, alors même que dans certaines circonstances elles pourraient être bonnes. C'est l'exemple de l'affaire Cambridge Analytica, qui vient du fait que l'API de Facebook est très

ouverte et que l'on peut en tirer ce que l'on veut. Concrètement, c'est très utilitariste, conséquentialiste et bénéfique pour beaucoup de gens, notamment pour des entreprises ou pour des médias qui veulent avoir des données plus fines de leur audience, mais ça peut aussi avoir une énorme conséquence négative. Par exemple, Cambridge Analytica qui va exploiter des données à des fins politiques ou à des fins de manipulation des utilisateurs. Cette utilisation particulière justifie le fait que l'intégralité de l'API devrait être beaucoup plus strictement contrôlée qu'elle ne l'était auparavant. Ça, c'est une application déontologique de se dire que même si beaucoup de personnes pourraient bien utiliser l'API, et bien on va la restreindre fortement car dans certains cas non prévisibles ça peut être détourné et mal utilisé.

Y a-t-il des domaines où l'éthique n'est pas importante dans les algorithmes ?

Il y a évidemment des domaines de la vie qui sont moins sensibles que d'autres. Si un algorithme biaisé sur Netflix ne recommande que des séries d'une certaine nationalité, par exemple des produits culturels américains, c'est surement moins « grave » à court terme par rapport à un algorithme prédictif pour la police qui inciterait à se rendre dans certains types de quartiers ou dans la justice prédictive, car il y a moins de conséquences directes sur les humains.

Néanmoins, on peut se dire qu'il faut toujours réfléchir à deux fois avant de créer un système technique. Je pense que ce n'est pas tant le fait d' avoir des situations où l'on peut se passer complètement de l'éthique, ce sont plutôt des domaines où les entreprises ont atteint une telle taille, qu'elles ne peuvent clairement plus faire l'impasse sur ces questions. Il faut plutôt prendre la chose dans l'autre sens : il ne faut pas penser que certains peuvent se passer de l'éthique car il n'y aurait pas de conséquences directes, mais c'est plutôt ceux qui ont un impact considérable qui doivent y faire particulièrement attention. Si les autres ne le font pas, ça ne sera pas dramatique, mais c'est quand même mieux. Par exemple, je travaille pour une entreprise qui crée des logiciels pour des bibliothèques, des musées et des centres d'archives, ce n'est donc pas à un niveau de criticité très élevé. Cependant, quand les institutions publiques pour lesquelles on travaille nous demandent d'envoyer les fichiers utilisateurs pour faire des analyses statistiques, on veille à avoir un respect du RGPD notamment, car on sait que l'on peut exploiter les données. Avec les données que l'on stocke, une collectivité locale pourrait connaître la rentabilité ou la performance au travail des employés d'une bibliothèque ou d'un musée. Directement, il n'y a pas d'impact, mais indirectement, si on ne fait pas attention, il peut y en avoir.

Selon vous, qui doit décider de l'éthique à mettre en place dans les systèmes algorithmiques ?

L'exemple qui est assez intéressant de ce point de vue là, qui est tiré des formes d'éthiques appliquées, c'est celui de la bioéthique. La différence entre l'éthique médicale et la bioéthique est que le médecin respecte son serment d'Hippocrate avec un patient spécifique, alors qu'en bioéthique, il s'agit de savoir si on autorise ou pas, et dans quelles conditions, certains sujets comme l'euthanasie, ou auparavant l'avortement par exemple. Dans tous ces sujets de bioéthique, ce ne sont pas que les médecins qui décident : on demande aussi l'avis de personnes qui travaillent dans d'autres domaines et qui peuvent éclairer la décision à prendre tels que des philosophes, des historiens, des sociologues et même des théologiens ou des personnes de la société civile. On peut aussi aller demander l'avis de la population à travers des consultations X ou Y. Cela s'est justifié par le fait que les changements qui seront opérés dans la déontologie des médecins, déontologie qui est normalement interne à leur profession, aura un tel impact sur la population que cela justifie qu'elle ait son mot à dire. Je pense que dans le

domaine de la technologie, à l'image de le bioéthique, on pourrait peut-être parler de technoéthique.

On devrait, de la même manière, avoir des personnes représentatives de la société civile qui donnent leurs avis sur ces questions-là. C'est en parti le cas, car on a déjà des structures telles que le CNUM, la CNIL, l'ARCEP, etc. Elles vont mettre en place une forme de régulation de certains aspects des technologies, ou vont fournir des avis sur ces questions de prospectives. Cependant, c'est généralement très juridique, notamment dans le cas de la CNIL, ou sans contrainte d'application, notamment avec le CNUM. Tout l'enjeu est dans l'indépendance de ces structures. Ça me rappelle le cas ou Google avait mis en place un comité éthique global à l'entreprise : le comité a dû démissionner au bout de quelques jours seulement car il n'était en réalité pas si neutre que ça. S'il n'y a pas de neutralité, alors ça ne sert pas à grand chose.

Au sein de votre expérience professionnelle, avez-vous un exemple d'un challenge éthique lié aux algorithmes auquel vous avez fait face ?

Oui, ça arrive assez régulièrement, et c'est arrivé aujourd'hui d'ailleurs. Ce ne sont pas des enjeux éthiques énormes mais ce sont des questions que l'on se pose à l'échelle individuelle ou de l'association chez Designers Éthiques. En ce moment, on communique sur l'évènement *Ethics by Design* qui doit avoir lieu en septembre ; pour communiquer, on utilise notamment des newsletters. À chaque fois que l'on organise un évènement, on collecte les adresses email des personnes qui participent afin de leur envoyer les informations pratiques et, à la fin, on supprime tous ces emails : on ne constitue jamais de fichier central des personnes qui font partie de notre réseau. Sauf que ça, en termes de communication pour les évènements que l'on organise, c'est nul, car l'on repart à chaque fois de zéro sur notre capacité à toucher une communauté.

De plus, de manière très pragmatique, ne pas être capable de vendre toutes les places pour un évènement peut mettre en péril la viabilité économique de l'association. On se pose donc très régulièrement ces questions : faut-il arrêter de collecter des adresses email pour garder cette approche moralement vertueuse de rendre totalement anonyme toute relation avec l'association, ou faut-il que l'on constitue un fichier de contacts, ce qui sera moins vertueux mais qui permettrait à l'association de grandir et d'avoir une action plus efficace ?

Nous arrivons aux questions finales d'ouverture. Quel est votre point de vue sur la place des biais dans les algorithmes pour la société à long terme ?

Honnêtement, j'essaye de ne rentrer ni dans des scénarios ni utopiques, ni dystopiques pour le futur. Je pense que cela va plutôt continuer comme c'est le cas actuellement, au sens ou la place des GAFAM va s'accentuer dans le numérique. On va probablement rester dans une situation assez similaire, avec un numérique américain et un numérique chinois qui sont très omniprésents et qui exploitent beaucoup les données personnelles, ainsi qu'un numérique européen qui essaye d'avoir une vision plus de défense des droits individuels. Cependant, je ne suis pas vraiment optimiste sur le fait qu'un jour, on arrive à une situation où tous les citoyens seraient sûr à 100% que tous les services techniques et numériques qu'ils utilisent soient complètement dépourvus de biais et totalement éthiques, dans le sens du langage courant. Je ne pense pas non plus que l'on arrive un jour à une situation totalement dystopique comme on peut voir dans des fictions tel que Black Mirror.

On parle souvent de l'équilibre entre humains et machines à mettre en place dans le processus de décision, quel est votre point de vue ?

C'est intéressant car c'est une question qui se pose depuis le tout début de l'histoire de l'informatique, notamment entre Engelbart, le chercheur de Stanford qui a inventé la souris et l'interface graphique, et Licklider, qui travaillait à l'ARPA, l'agence américaine de Défense née pendant la guerre froide ayant créé Arpanet, l'ancêtre d'Internet. Ils avaient deux visions fondamentalement opposées de ce à quoi devaient servir le numérique. D'un côté, Licklider pensait que les humains auraient beaucoup de tâches en moins à faire et que les machines allaient devenir autonomes et produire pour les humains, de l'autre côté, Engelbart pensait que le numérique allait pouvoir encapaciter l'humain. Personnellement, je pense plutôt comme Engelbart, le numérique peut encapaciter l'humain, mais je ne suis pas sûr que le numérique, ni la technologie en général, remplacent l'Homme un jour. On dit souvent que la technologie va remplacer l'humain en prenant l'exemple des caissiers dans les supermarchés qui disparaissent, c'est vrai en partie mais d'autre part, il y a plus d'ingénieurs recrutés pour fabriquer des caisses automatiques. Certes, l'impact de la technologie peut être très fort sur certains métiers spécifiques, mais je ne suis pas sûr qu'elle remplace le travail que doit faire l'humain en permanence.

Quels sont les grands enjeux actuels ou à venir de la recherche concernant l'éthique des algorithmes ?

Je dirai qu'il y a notamment le sujet de la 5G et celui de la reconnaissance faciale, qui sont les deux grands sujets technologiques du moment et qui englobent beaucoup d'enjeux éthiques. La 5G est presque un choix politique, voire de société, car on dit qu'elle engendrera davantage de connexions, d'objets connectés, etc. Cela sous-entend de repenser tous nos réseaux, avec une explosion du nombre d'objets technologiques, ce qui aura un impact écologique important. Concernant le sujet de la reconnaissance faciale, il pose énormément de questions sur les libertés individuelles et sur la vie privée. Je pense que ce sont les deux sujets principaux, et ils rejoignent le sujet majeur et global de l'environnement et de l'écologie, qu'on peut aussi appliquer au numérique. La 5G est une facette de cette question, que l'on va par exemple pouvoir retrouver dans la *low-tech*.