Artificial Intelligence, Data, Ethics: An Holistic Approach for Risks and Regulation

Alexis Bogroff, Dominique Guegan

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers

CES
Centre d'Économie de la Sorbonne
UMR 8174

Artificial Intelligence, Data, Ethics: An Holistic Approach for Risks and Regulation

Alexis BOGROFF, Dominique GUEGAN

2019.12

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
https://centredeconomiesorbonne.univ-paris1.fr/
ISSN : 1955-611X

Artificial Intelligence, Data, Ethics
An Holistic Approach for Risks and Regulation

Alexis Bogroff[*], Dominique Guégan[†]

June 24, 2019

Abstract

An extensive list of risks relative to big data frameworks and their use through models of artificial intelligence is provided along with measurements and implementable solutions. Bias, interpretability and ethics are studied in depth, with several interpretations from the point of view of developers, companies and regulators. Reflexions suggest that fragmented frameworks increase the risks of models misspecification, opacity and bias in the result. Domain experts and statisticians need to be involved in the whole process as the business objective must drive each decision from the data extraction step to the final activatable prediction. We propose an holistic and original approach to take into account the risks encountered all along the implementation of systems using artificial intelligence from the choice of the data and the selection of the algorithm, to the decision making [1].

Keywords: Artificial Intelligence - Bias - Big Data - Ethics - Governance - Interpretability - Regulation - Risk

JEL classification: C4 - C5 - C6 - C8 - D8 - G28 - G38 - K2

1 Introduction

For forty years, regulators have been trying to build a uniform setting to compute the risks associated with the activity of the banks and the insurance companies (e.g. Bâle III and Solvency II). The philosophy developed during this period - for market, credit, operational risks - in order to compute a capital requirement for the banks, was to impose (i) a percentage of the capital, or (ii) a computational procedure based on an internal modeling developed under specific assumptions and specific statitical tools, privileging the VaR and ES measures. After the 2007 - 2011 crisis, banks and insurance companies focused on strengthening the compliance and risk management departments, encouraged by regulatory developments and a strengthened financial supervision. This trend has stabilized, and the stakes are now relative to data and algorithms.

Financial regulations tend to use a single analytic framework, disregarding the decision-maker's situation. This can lead to transposing tools for some situations where they are not relevant. The

[*]University Paris1 Panthéon-Sorbonne; mail: alexis.bogroff@univ-paris1.fr

[†]University Paris1 Panthéon-Sorbonne labEx ReFi France and, University Ca'Foscari, Venezia, Italy; mail: dguegan@univ-paris1.fr.

[1]This work was achieved through the Laboratory of Excellence on Financial Regulation supported by PRES HeSam under the reference ANR-10-LABEX-0095

uniqueness of the mathematical formalism can create mistakes as soon as it is transposed from one domain onto another, distorting our intuitions and damaging our understanding of risks. To prevent this problem from being reproduced in this new context, the risks related to systems of big data and artificial intelligence should be carefully analyzed internally as soon as these technologies are used in a company, although the applications only represent a small part of the business activities.

The concept of big data has been largely investigated in relation with its uses in diverse applications, concerning the technical approach of storage and the use of specific and incomplete data sets. These large data sets are now used for modelling, analyzing specific features, discovering new patterns, linking events and yielding predictions through models that are commonly called "machine learning". The last concept is not new and was introduced in the 1950s. Nonetheless, its fatuation is recent and is partly due to the capability of new technologies, more efficient to catch, store and organize these big data sets, and to the simplification of such processes and their availability through cloud services.

Although it enables the development of many tools with the potential of bringing good to the society, it might also generate or inflate risks. Thus, many questions are raised for regulators with the rise of new concepts like social networks, connected objects, the increase of information technologies, the advances in data science, and more specifically the development of deep learning models (associated with their notions of bias, interpretability, auditability, etc.), the spread of dynamic architectures handling continuous data flows, the emergence of the data scientists status, etc. Furthermore, relatively recently, privacy regulations along with a set of compliance rules were introduced, temporally increasing the instability of the economic system.

Thus, what are the good scenarii at the regulatory level? What are the political trade-offs between prudential issues and other economic, macro-financial and social issues? These new dimensions increase the complexity of some risks that are more transversal by nature, and hence hard to classify. Finding solutions, measuring and regulating such risks is therefore challenging. The reflexions on the integration of these technologies in the decision processes of companies should be thought and implemented quickly, since methods and frameworks change drastically in this fast-pace environment, yet the current proposals on the different facets of this problem are very diversified.

The paper strives to depict as precisely as possible this complexity in order to help managers in implementing thoughtful solutions and ease the task of imposing new standards for the regulators. We thus analyze the potential risks associated to the algorithms used for making decisions, and illustrate methods to avoid confusions due to their complexity and lack of transparency. We also investigate data sets and the methodologies that both enable to assess their quality and legal status. Indeed, can regulation clarify data holders' duties in order to answer to several questions considering their storage politics, analysis and transformations of the data, supply chain in real time, access rights, privacy and ethics, governance, and audit?

The objective of the paper is to provide a uniform risk management framework for these new technologies that could lead to coherent strategies supporting their development without hindering them. A study of the risk panel is therefore crucial to better assess the specific needs for regulation.

Our proposal is original in its approach by considering the whole process inseparable, and thus studies the data selection, their pre-processing, the model selection, its training and interpretation a nested way. In other words, we consider that any part of the process analyzed independently would lead to erroneous specifications, misinterpretations and bias.

This work is not a review article. It as an oriented analysis of the risks arising throughout the imple-

mentation of infrastructures of big data and artificial intelligence, considering in parallel a technical and a legal point of view. The subject is large and this paper is not exhaustive. It lies on fundamental and recent researches and applications, for which some references are suggested at the end of each section.

The paper first introduces briefly the concepts of big data and artificial intelligence in Section 2, then studies their risks in section 3 and 4 respectively and provides some straightforward solutions. The problematic of bias is then developed along with the current regulation in Section 5. Recent and more technical developments associated to the interpretability of models and the fairness of their decisions are discussed in Section 6. Section 7 concludes.

2 Presentation of the concept of big data and artificial intelligence

2.1 Big Data

The collect of Data begins in the XVIIth century, more or less around 1662 with the publication of Graunt'book Graunt (1662) providing some indications on the population's mortality in London. Then, in the XVIIIth century, a scathing clash breaks out between Bernoulli and d'Alembert on the use of data: Bernoulli uses a statistical approach and favors global improvements for the population, while d'Alembert priorizes the fate of each individual. D'Alembert could not accept a method that would improve the average population's condition at the cost of some people suffering from aggravating conditions. Then the statistical interest for data was more developed since the end of the XIXth century, and the arrival of big data at the beginning of the XXIth century transforms the era of quantification to datafication. The recent phase can be illustrated by the development of more personalized scores like for insurance contracts, probably not anymore based on solidarity values in the future, and with the profiling methods used in case of recidivism in US, dictating a person's possible release.

If small tables are handwritten at the beginning, then typewritten, and transformed into worksheets with the use of computers at the end of the XXth century, the big data paradigm greatly differs as it concerns Terabytes (TB) and Petabytes (PB) data sets. During the 90s, the rise of data mining exploration can be observed, exploiting large data sets to discover more complex patterns and using high performance algorithms running efficiently through parallelization. Around 2010, a paradigm shift takes place with the arrival of big data sets (PB, ZB) considering that a big amount of data with heterogeneous, diverse and complex sources, due to autonomous origins, must be treated with distributed and decentralized control. Thus, for 10 years, a new industry is emerging with the arrival of platforms handling data storage and analyzing the quality of the data (which is a key point for risk measurement), the introduction of privacy regulations with a set of compliance rules, the development of deep reinforcement and transfer learning models, associated to the notions of bias, interpretability, auditability, etc., the integration of dynamics inside any process to integrate continuous data flows, the emergence of the data scientists status, etc. These phenomena are very important and have both bright and dark sides.

From the regulatory side, the arrival of big data strategies threatens the latest trends in financial regulation related to the simplification of models and the enhancement of the comparability of approaches chosen by various entities due to a new complexification of the concepts and of data sets (with data lakes). Thus, the questions around the necessity to understand, control and use these new concepts within a single framework are opened. New alternatives in terms of risk measures and risk management need to be developed.

In another hand, big data is also a cultural phenomenon. From the technology side, the development of algorithms specifically designed for big data sets permits to optimize computations, storage, analysis, linkage and comparisons (a lot of tasks that a human would not do so quickly), and from the analytic side, global models permit to identify patterns for economic, social, technical, and legal reasons. However, the big data phenomena might be overestimated by certain for whom it provides a higher form of intelligence and knowledge with an aura of truth, objectivity, and accuracy. This clearly needs to be carefully re-considered and opens a philosophical debate. Indeed, on one hand big data triggers both utopian and dystopian rhetorics, as it is a powerful tool to address various societal ills (health, security, climate), but on the other hand it can also be seen as a manifestation of Big Brother with invasions of privacy and reduced civil freedoms. The nuanced and subtle shifts that are underway should be examined carefully.

As soon as an enterprise decides to use big data as an innovative process, the objectives, infrastructure and legal aspects need to be specified: (i) Why is big data interesting? (ii) Why are financial institutions willing to evolve within a big data environment? (iii) Do they really need it? (V) What kind of volumes are at stake? (Vi) What types of data are considered? (vii) Which platform is required? Is it possible to build its own platform? (viii) Legally speaking, who is the owner of the data? (ix) What type of data can be used? (x) To which extent is scrapping allowed? (xi) What is the current legal and regulatory environment? What are the risks associated to this use?

The interest of big data mainly lies within the new applications that it enables. These data are valuable information about the activity of individuals in areas as diverse as marketing (to better understand customers' needs), commercial (new tools to improve the relation with the client, to increase the offerings and services), security (to increase the detection of money launderings), financial (to decrease the costs, optimize portfolios), and also for health, medical records, genomics, chemical process, biological science, life sciences, transportation, geolocation, astronomy, atmospheric science, education, sustainable development, natural disaster, physics, industry, resource management, private or public sectors, military surveillance, social networks, text, photography, audio, video, etc.

The market sees big data as a pure opportunity: marketers use it to target advertising, insurance providers to optimize their offerings and bankers to read the market and optimize portfolios. This raises several challenges concerning the models used to treat big data sets and their associated risks. These new technologies seem to intrigue the leaders with a mixture of doubt and seduction. The management of this massive data, and more generally the data science, is part of the will of the companies to enhance their customers' knowledge, in order to anticipate their demand and design better products and services that could meet these needs. In short, this is a new approach to the all-round digitization, which seems to address the need for organizations to respond as quickly as possible to market demands, ideally in real time, or to even predict the risks. The stakes for companies are huge and will require a complete overview of their management systems. Hence the need to master the associated concepts and risks they originate.

In fine, we can imagine a methodology for building a safe system for big data architectures, as being one implementing the following steps:

1. Tier I: To build a platform updated continuously
2. Tier II: To define a process controlling the origin and quality of the data, while preventing the introduction of bias in the process
3. Tier III: To explore the data and define specific objectives
4. Tier IV: To develop adapted complex modellings with their interpretability

5. Tier V: To develop a risk management framework with adequate risk measures

6. Tier VI: To provide final decisions associated to the use of the big data set in relation with the modellings

2.2 Artificial Intelligence

The research field of Artificial Intelligence (AI) can be considered as starting at a workshop at Dartmouth College in 1956 with Allen Newell and Herbert Simon declaring ”machines will be capable, within twenty years, of doing any work a man can do”. John McCarthy, Marvin Minsky, Arthur Samuel (IBM) then become the founders and leaders of AI research, producing programs that can be described as ”astonishing” (dividing in small pieces the objectives to attain): computers are winning at checkers, solving word problems in algebra, proving logical theorems and speaking English. At the same time Turing (1950) is trying to answer to the following question: ”Can a digital computer take the place of a human being in the game of imitation?”. At this period the research is based on a logical approach. In 1960, the research in the U.S. is heavily funded by the Department of Defense, and in 1980 begins the development of expert systems.

At that time, the focus is on intelligence exhibited by machines with autonomous program, being able to answer different kinds of questions. A machine tries to mimic ”cognitive” functions that humans associate to other human minds, such as ”learning”. The speech recognition is now a standard task, but a human-like comprehension of the language is not yet reached. Thus, robots are able to move, take objects and in some ways understand language. To achieve these tasks, machines are ”trained” by computing processes that permit to converge toward a desired objective by experimenting the data.

At the end of the twentyiest century, this logicist approach is replaced by the neuronal approach, imitating neuronal biological processes. Advanced statistical techniques (loosely known as deep learning) permit to treat large amounts of data using faster computers, enabling advances in machine learning and perception. By the mid 2010s, machine learning applications are used throughout the world. Thus, the field of AI experiences a crucial development, due to the conjunction of serveral factors: (i) more powerful and dense processors (GPU), (ii) lower storage cost, (iii) easy access of cloud computing, (iv) large data set available, (v) available AI platforms like Microsoft, Amazon, Google, DataRobt, etc.

One main aspect of AI is to create algorithms that discover new patterns and new behaviors from different data sources. Two types of training are conducted: (i) supervised learning: to classify data according to their labels provided by humans (e.g. spam or not spam) (ii) unsupervised learning: to create groups using unlabeled data. The resulting classes thus greatly depend on the method used (e.g. blue circles and red crosses could equally be grouped by their color or their shape). Some hybrid models combine both approaches into a semi-supervised framework in order to leverage their capacities by using a supervised learner while enriching the training set with unlabeled data.

Challenges in AI endlessly become harder with the intent to continuously generate more global models: starting by combining locally discovered patterns to form a unified view, then, by analyzing model dependencies between distributed sites and fusing decisions from multiple sources to gain a best model out of the big data, and finally, by creating a big data mining framework able to consider complex relationships between samples, models, and data sources with their evolving changes along time and other possible factors. The futur carrying out concerns, for one side, unstructured data which can be linked through their complex relationships to form useful patterns, and on another side, the use of growth of data volumes and relationships.

AI uses different methodologies depending on the complexity and type of application. For simple problems, the use of basic models such as Trees, Random Forests (RF), k-Nearest Neighbors (k-NN) and Support Vector Machines (SVM) might be adapted. More complex models like neural networks (multi-layer perceptrons), genetic algorithms and boosting methods can also make the standard models more powerful, often at the cost of some opacity (randomness, averaging of multiple simple models, etc.). Some of the most complex tasks (computer vision, text translation) are standardly done using deep models, with a competitive accuracy, but at the cost of an even greater opacity. Then, Deep Learning (DL) was introduced as architectures enabling to create internal representations useful to solve complex problems. The categories forming machine learning could be visualized as follows:

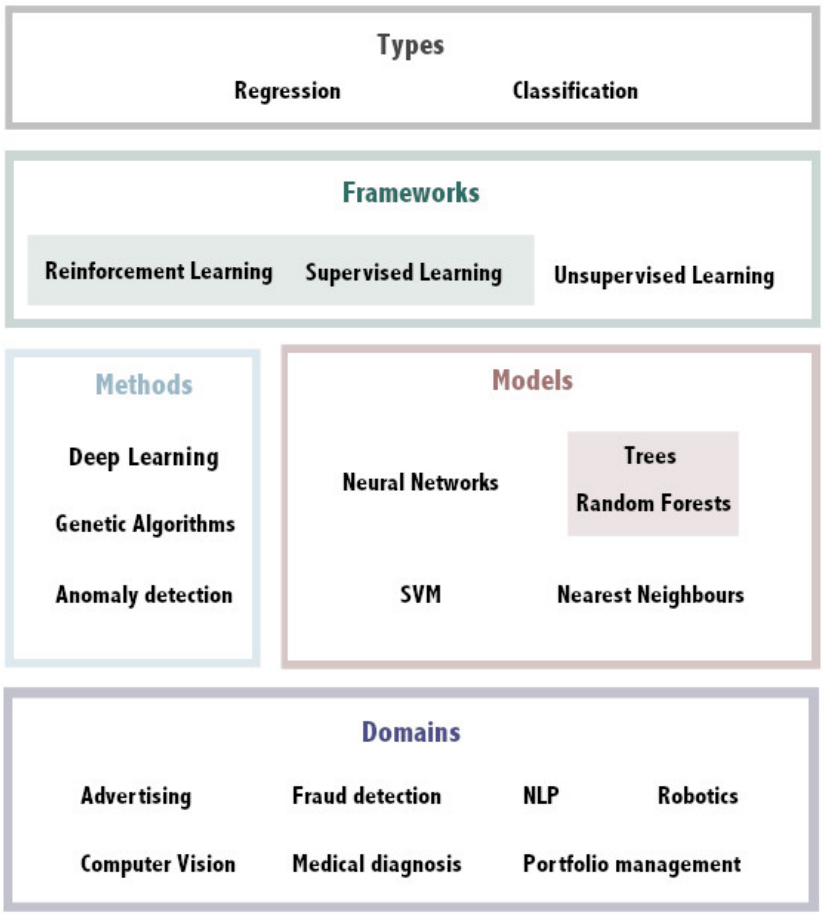

Figure 1: A representation of machine learning categories.

AI, as we have just seen, is a set of theories and techniques developing complex computer programs capable of simulating certain traits of the human intelligence. The widespread supervised learning models can be formalized as follows: a target function f maps an input variable X to an output variable $\hat{Y}$: $\hat{Y} = f(X)$, with Y the objective (ground truth). The mapping f which constitutes the model can be more or less complex. A graph can depict the process as follows:

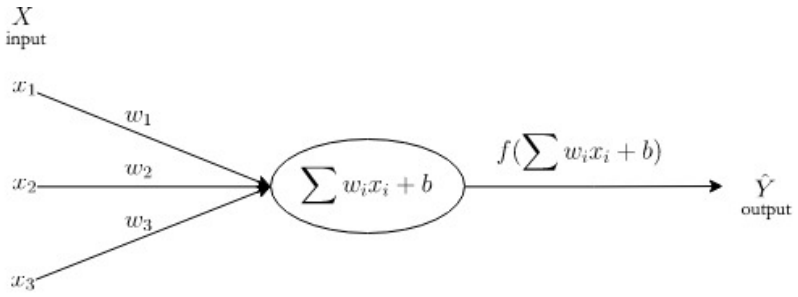

Figure 2: A simple representation of machine learning model with a single neuron.

AI enables humans to be released from continously more complex but repetitive tasks, and fasten processes on elaborate diagrams that would otherwise take a lot of time to humans. The machines will never become a person, and the person will never become a machine either. Both are complementary and a cohesion is favorable to the development of a company. Nevertheless, the tools that are developed within the AI framework can generate risks that are important to be aware of, to measure and to control.

2.3 Risks identified for the use of big data and AI and some references marks

What are the risks associated to big data and how can we control them? Are the risks different depending on the domains in which we use the large data sets? Some risks are inherent to the data sets (poor quality, representativity, etc.), while others depend on the algorithm (model selection, training and parameterization, etc.), others belong to the subject itself (covariance shift, unpredictability, etc.) and finally to the process, for instance operational risks (cyber, conduct, IT risks) and compliance risk (the regulator mainly focuses on privacy). For example, considering financial applications, the question of compliance appears fundamental because of auditing. It is also crucial to understand the decisions as consequences can be important, at least financially. A great list of risks are also specific to the finance industry: market risks, market manipulations, systemic risks, diverse bias coming from data and algorithms, and also extreme risks.

We analyze some forms of these risks in the following in order to provide solutions and strategies to measure and to avoid them. The objective of the paper is to define a unified risk management framework for these new risks. Thus, a study of the risk panel is important to assess the need for regulations, control, and to find coherent measures that could support the development of these technologies without hindering them.

We provide now some references for the topics introduced in this Section. The guidelines on risk measures can be found in Basel-I (2004), Basel-II (2016), Basel-III (2017). The research literature on these subjects is large, we suggest Jouini et al. (2004) and Guégan and Hassani (2019) where we can find a lot of references therein. Some historical references concerning the evolution of the use of the data are Quetelet (1846), Galton et al. (1909), Mayer-Schönberger and Cukier (2013) and for a recent review Barry (2019). The domain of storage, properties of the data is investigated in many papers, we can cite Xing et al. (2015), Demchenko et al. (2014) among others. For Artificial intelligence, some fondamentals are in Simon (1956) and Rosenblatt (1960). More recent presentations are done in James et al. (2013) and Russell and Norvig (2016). The development of algorithms designed to handle huge volumes are presented for instance in Al-Jarrah et al. (2015), and Goodfellow et al. (2016). We have cited a lot of research domains interested by artificial intelligence, some references are Ding

et al. (2015) for financial predictions, but also Renault (2017) and Feng et al. (2018), and Nevmyvaka et al. (2006) for trading optimisation with reinforcement learning models. Other applications are credit scoring, for instance with the works of Addo et al. (2018), Li et al. (2017), Banasik and Crook (2005), and also collusion strategies through reinforcement learning Calvano et al. (2018), Klein (2019) or OECD (2017). In social sciences, some applications concern human geography (Kitchin (2013)), and also recidivism prediction (G'Sell (2018)), among many others.

3 Risks associated to the use of Data

The pre-processing and analysis of the data appear as the main tasks when working with machine learning algorithms. Data are indeed determinant in the results yielded by the algorithms and therefore in the decisions taken and it is considered that 80% of a data scientist workload consists in analyzing and preparing the data. Although the subject is not new for professionals managing data bases and distributed platforms, a discussion is proposed on several kinds of risks associated to the notion of big data, including data quality, storage, pre-processing, external risks, privacy and ethics. Section 6.1 provides more solutions for the data sets representativity and legal problems with a higher technical point of view.

3.1 Data Quality and Complexity

The data have several properties which can be classified into two categories: (i) their quality, i.e. the source, coherence and representativeness of an information set, and (ii) their complexity, i.e. the dimension, generating law and balance of the labels.

1. The data quality has two major aspects: the technical neatness of data samples (the absence of NAs[2], a constant formatting of each variable, the absence of outliers, etc.) and the statistical representativeness of the data set. It is worth noting that from the point of view of a company using data from a provider, the source could have two minings: it could concern the provider as it is the source of a data flow, but could also concern the object from which the data is generated, also called the underlying or origin of the data.

   The technical neatness of data can be analyzed as an inner property of a data set. In order to control the coherence of data samples, one must analyze each of its components carefully. Indeed, NAs not necessarily represent mistakes since the missing of a value can actually be the information (although it should then be encoded differently to be processable by the models). Detecting outliers suffer from the same issue since extreme data may not be an outlier. The quality regarding these components is thus hard to qualify without having a sufficient knowledge of each variable. Nevertheless, some properties are easier to evaluate like the format of the variables.

   The representativeness of the information set requires to be aware of a clear objective to enable its assessement. Indeed, a same object-phenomenon can be represented in several ways, hence, the data generated then depends on the objective of the model thereby defining the interest of a specific representation. Thus, as a provider might not be aware of the future uses of its data by clients, a reasonable way to ensure the data quality could be to aggregate meta-data that would describe how the data were extracted and to which phenomenon they correspond to. For example, if 98% of Facebook publications are posted by only 4% of its users, such data can coherently be exploited to understand what users read on the media. On the contrary,

[2]NA: non available, the use of this terminology is widespread

using these data cannot provide any robust information on the type of publication posted by an average user. Indeed, it only represents what 4% of users do publish, and the other 96% users are not represented. Knowing the distribution of users' publications in this example could prevent from erroneous uses of the variable. It is thus possible to measure the data quality to a certain extent by the meta-information provided (source, extraction method, distribution of contextual variables, etc.).

Furthermore, some data sets only represent a portion of a whole information set. This issue should be considered carefully as algorithms going into production can fail rapidly in case of an inadequation of the trained model with out-of-sample data. The subsets are statistically representative if they have the same distribution than the latter. However, depending on the type of task and availability of data, estimating the risks of failure of an algorithm in production can be challenging. Survey and panel methods can help in constructing robust subsets of data.

2. The data also have inner properties predetermining the complexity of their use and processing: the dimension, stability and balanced labels, as described in the following:

   (a) Dimensionnality. It is a crucial aspect of a data set which makes some problems more elitists than other. For instance, the computation of million high-quality images or billion texts data bases are so expensive that merely few global companies are currently able to properly tackle. On the opposite, applying machine learning methods on character recognition and client recommandation tasks is more accessible. The dimension can be measured by the weight of a sample and of the whole data set in a first approach. Fortunately, advances in computer science enable the improvement of such trainings using parallelization accross multiple servers, on specific hardware GPUs or TPUs, and exploiting the sparsity of these data sets. These progresses also partly explain the development of Deep Learning methods (with the explosion of data volumes). Also, these technologies tend to be open-sourced and some platforms provide free computation on GPUs and even TPUs for a specific use (tipically for researches and experimentations).

   (b) Stability. A deep understanding of the origin of the data (i.e. the underlying object from which the data is generated), is also fundamental for a company in order to maintain a control on its results. Some underlying objects have unstable distributions that could not be understood in the same manner than simpler objects, as some could be naturaly shifting in covariance or exploding, for instance financial asset prices are less stable than human body measures. On the long run, extreme events would probably transform the original distribution of these complexe unstable objects. More tricky objects could show a stable distribution on a long period until the occurence of a dramatic shift. For instance, data from a social media could evolve drastically as some communities move from a social media to another (because of the evolution of their habbits, or because of new laws affecting the users' experience), hence the importance to continuously inspect the coherence of these data flows. Furthermore, the complexity of inspection increases with the number of sources required to obtain the final aggregated data flow.

   (c) Unbalanced data[3] generally increase the difficulty of problem solving since learning a distribution from few examples is often a hard task (if the distribution is complex), therefore corresponding to more complex data structures. More developments are done in other sections.

[3] Unbalanced data sets: for which the proportion of some classes is greatly superior to others

3.2 Storage

When using big data sets (e.g. peta bytes) an important step consists in the storage of the data. It creates several technical challenges as it requires efficiency, flexibility and security. The problem can be partly addressed by third party operators that store data on their "cloud servers" (or platforms) and propose a diverse offer that could fit many kind of uses. However the distant system increases the number of entry points for cyber-criminals which could then stole information or modify data. This issue might be reduced by diminishing the number of intermediaries.

The use of major economic actors (providing platforms) increases the risk of unbalanced commercial relations which could lead to artificially increased prices or limited services. It might also be a lever for countries hosting such companies. Therefore, enterprises can adopt several strategies: to create their own platform, to use the platform of a provider, or to implement a mixed architecture.

- When an enterprise uses the cloud or a platform:
  - it is a flexible way to treat the data for the enterprise as it can rent the storage capacity to fit its needs. The use of a provider permits interoperability, strategic governance, up to date technologies, robust security systems, and the data server being remote can be shared. Furthermore, large scale platforms enable the optimization of the costs thanks to mutualisation. The storage of the data can be more secure because most platforms have several copies of the data to avoid losses. Companies can also use technologies that they are unable to develop.
  - however, several risks emerge: cyber-risk, governance risk (as it shares the information with the provider, requires the comprehension of the provider's system and creates a technological dependence), security risk (how to be sure of the security of this specific platform?), risk of data loss as the data are not inside the enterprise (platform can be hacked, disappear, etc.). Also, the relation with the provider requires trust in this third party and detailed contracts taking into account all the facets of their interaction. From a macroeconomic approach, a governance and a systemic risk can be emphasized when data of several companies are centralized and shared on a unique platform, which also increases a reputational risk.[4].
- When an enterprise wants to create its own data lake, it is necessary to have an interdisciplinary team with statisticians, computer scientists, DPO[5] for the legal questions, and domain experts. All these persons are involved in the process of the source, the quality, the compliance of the data, but also the security processes and the use of the algorithms for the treatments of the data with respect to the objectives fixed by the enterprise.
  - the expertise thus belongs to the company that does not depend on external requirements in this regard. It thus increases the control of the whole process.
  - this strategy has a high cost and will be probably choosen by big firms as it requires to invest in human and technological capital.

3.3 Data pre-processing

It is known that data pre-processing represents the major work of a data scientist workload. Transformations can consist in several repetitive tasks like harmonizing formats, detecting and deleting outliers. Making such decisions is a highly sensible task as developed in the following.

[4]Clipeum is a new development on the storage of the data adopted by certain banks in order to mutualize and share information to better prevent money laundering.

[5]DPO: originally Digital Protection Officer and shifted to Data Protection Officier

1. Presence of outliers. Determining the presence of outliers is a hard task that requires an in-depth knowledge of the studied object. Indeed, keeping outliers could hinder the learning of the algorithm, but removing information also risks to deform the real data distribution and eventually preclude the algorithm to acknowledge extreme cases. Even though outliers necessarily lay far from the rest of the distribution, they cannot be detected on this sole information. Put another way, extreme cases can be some of the most important informations of a distribution and their deletion would mechanically lead to underestimating the variance and kurtosis of the phenomenon, which are fundamental in critical applications potentially impacting humans. Thus, a domain expertise is required for the process of deleting outliers since statistical informations are not sufficient for their identification (at least before a robust anomaly detection algorithm is trained on other data of the same domain). Similarly, normalizing data could on one side improve the learning efficiency and, on another, bribe information as it forces the distribution to follow a strong hypothesis.

2. NA entry. Concerning the imputation of NAs, this can be done using several methods: imputing the average value of the variable or copying the previous or the next data sample, which could make sense if there are important local correlations. These methods greatly depend on the object that generated the data and no convention could hold for any object. In the worst case scenario, the NAs are not mistakes but inversely represent the information by their absence. For instance, in a form, the absence of an answer in some sections could indicate that the person agreed (e.g. fill-in to make a claim) or disagreed (e.g. check the box for allowing the use of your mail for marketing purposes) with the question asked. Hence, the imputation of such NAs would lead to data destruction. However, since NAs break most of the learning models, these data should in any case be replaced by non NA values, but being careful to avoid merging NAs that represent information with NAs that are mistakes. Meta-data are thus required for the NA imputation process.

Several operations can be performed on the data between their extraction and their use as an input for the models. First, these data can be stored, whether internally or on an external platform. Then, backups can be created and some data are moved from a segment of the database to another. Finally, a large number of transformations can be applied. To ensure the integrity of data during these processes, it is possible to apply best practices known under the accronym "ACID" for: Atomicity, Consistency, Isolation and Durability. An operation satisfying these properties is considered a transaction. This avoids incomplete operations, ensures the respect of constraints, prevents from conflicting operations and losses from hardware failures. However, this does not ensure any compatibility of data with the model, nor does it prevent from biases. Isolation is the typical property that must be satisfied as soon as multiple operations often occur on the same data. Standard implementations easily prevent conflicts by locking processing data, but as speed requirements grow, the rules governing data operations should rather be based on "optimistic" or hybrid methods.

3.4 External Risks

We identify two risks and several others might depend on the domain of application.

- Provider's risk. Some risks could depend on the ability of a provider to remain constant: on its methods used to extract data and in the formatting of its variables. Any change in these categories could induce large costs for the client using the data depending on its infrastructure and ability to detect changes, and to adapt. The provider should thus limit these changes in order to be reliable. In addition to these technical considerations, a provider could also transfer its own risk as a company related to financial issues and involvement in illegal or unethical activities. A more specific risk could come from the extraction of data that is not allowed in the

country of the client. These aspects should be considered closely by the client when choosing its provider since it could result in an interuption of the data flow, have regulatory repercussions or hinder the image of the company.

- Manipulation of the source. Cyber-attacks or specific malicious interventions on data could impede, mislead or destroy algorithms performances. Most notable attacks introduce falcified or misleading data. More advanced methods from financial markets like those using high-frequency trading (HFT) could directly impact the sources of data to manipulate the competitors' algorithms.

3.5 Regulation on property, privacy and ethical issues

Although it seems that the collection of data cannot be fully regulated, can regulation clarify data holders' duties in order to answer to several questions regarding their storage politics, ethics, analysis of the data, supply chain in real time, data quality, transformations, access rights, governance, and audit?

1. Logs of data transformations. Since it has a critical influence on the results, any transformation on the data (operations, formulas, etc.) need to be logged and well documented for subsequent auditing purposes (whether the data scientists are internal to the company or they work at the data providers). Any missing information could be penalized during auditions, but as it represents an outsized amount of records, tailored methods could eventually be implemented to produce these online backups, otherwise the regulator could adapt to solely request the storing of the core actions. May a system similar to Git helps in keeping track of such transformations.

2. Privacy law. Practitioners need to know the laws on the property of the data in the different countries (RGPD in Europe, CNIL in France, etc.). It is necessary to ensure that the enterprise (or the provider) can use the data, even though the notion of big data is not defined in the law. Indeed, there is no right on data itself, but sui generis right exists regulating data extraction from databases: a "producer" of content (physical or moral person) bears the risks on the investment made to acquire, verify and present the contents. A company is considered to be the producer of a database if it is able to prove its investments regarding the collection and setting of the data, and nothing is said on the resulting owner of the patterns obtained after data processing. Thus, concerning extraction rights for the producer of data, he/she has a substancial, repeated and systematic right of extraction on its own database, and he can grant full or partial extraction rights to its co-contractors or any other party. Concerning the user he/she can extract data as long as it is not a substancial, repeated nor a systematic extraction. For instance, the period of data protection is 15 years but it is possible to extend the initial period by proving maintenance acts on the data set. It is hence possible to indefinitely extend the protection.[6] It already exists a penal risk (in France, 3 years in prison and 300.000 euros as long as the person does not respect the producer's rights).

3. Obligation rights. Another point, concerns the right to oblivion. It is only partially tackled with the right to obtain the deletion of (some) personal data, but indirect ways to identify a person are still under discussion, i.e. potentially not punished. Indeed, the data used to train a facial recognition model could be deleted, but the weights of the model would retain the identifying information.

    - A straightforward solution could be to refuse to be profiled by such advanced methods. However, this might not be possible in some states, and insurance companies would in-

[6] https://www.constellation.law/le-droit-du-big-data.html

terpret this request as a bad signal. If regulations could forbid to process this information, the way to apply and verify it is an open question.

- A more technical method could orthogonalize the result with the data to delete or to build new architectures that enable the deletion of information by design (which is not the case for a standard neural network).

Although these ethical risks increase with the development of big data, the control from authorities remains difficult because of the lack of a uniform regulatory framework. Privacy risks thus have to be analyzed with respect to the legislation of the country.

Some references concerning this Section are now provided. Discussions on big data, their properties and uses refer to Boyd and Crawford (2012). Various ethical issues are conducted in Nunan and Di Domenico (2013), Zwitter (2014), Landau (2015) and in Richards and King (2014), among others. Data privacy regarding their consequences is analysed in several papers in relation with the legal country rules, Ram et al. (2016) for China, and Goodman and Flaxman (2016) for Europe among others.

4 Questions and risks associated to the use of algorithms

In this part, we focus on the algorithms and investigate their characteristics and also some risks. What are the main problems businesses and regulators encounter with the use of algorithms? What are the risks they face: use of black box, forced calibration, estimation, choice of the targets, choice of the parameters, labellisation of the data in entry, verification of data sources, understanding of the processes?

What is an algorithm? "In mathematics and computer science, an algorithm is an unambiguous specification of how to solve a class of problems. Algorithms can perform calculation, data processing and automated reasoning tasks. Starting from an initial state and initial input, the instructions describe a computation that, when executed, proceeds through a finite number of well-defined successive states, producing "output" and terminating at a final ending state. The transition from one state to the next is not necessarily deterministic; some algorithms, known as randomized algorithms, incorporate random input"[7].

Thus, algorithms are a sequence of lines of code written by humans for a specific objective using a specific information set. Can an algorithm be biased, discriminating, neutral, fair, interpretable? These questions are addressed along with problems associated to their specific uses. Understandable models are sometimes called transparent, while incomprehensible models are called black boxes. However, what constitutes transparency? Will the algorithm converge? Does it produce a unique solution? Can answers be found in the architecture of the model or in its parameters? Could the complexity of a model be considered differently? Is it simple enough to be examined all at once by a human?

Considering the previous questions, two types of algorithms can be distinguished: those that make impactful decisions (e.g. credit assignment, recidivism predictions) and those that mostly evolve autonomously (e.g. surveillance, recommander systems for advertising). In the latter case, explanations are not necessary either because (i) there are no significant consequences for unacceptable results or (ii) the problem is sufficiently well-studied and validated in real applications, bringing trust in the

[7] Wikipedia

decisions of the system. In this section we investigate the former case, which decisions might have irreversible human consequences, falling within the predictive-prescriptive approach, and requiring an optimal control of the process and results. More precisely, recent questions around algorithms including interpretability, governance and operational risks are analyzed.

4.1 Interpretability, transparency and explicability

The new infatuation of interpretability of machine learning lies on two points. The first point concerns new regulations and the second the technical aspects of black box systems.

1. Interpretability and regulation. The European Union states that individuals affected by algorithmic decisions have a right to explanation. What form such an explanation might take remains an open question. The same regulations also suggest that algorithmic decisions should be contestable. So, in order for such explanations to be useful, it seems that they must (i) present clear reasoning and (ii) offer some natural way of contesting these propositions and to modify the decisions if they are inappropriate.

2. Interpretability and the technical approach of black boxes. The deployment of machine learning systems in complex applications has led to a surge of interest in systems optimized not only for expected task performance but also for other important criteria. Indeed, for such a system to be used safely, it is critical to also satisfy the following auxiliary criteria: safety, fairness or unbiasedness, privacy or right to explanation, reliability or robustness, causality and trust. Nonetheless, a popular fallback is the criterion of interpretability that is sacrificed for the accuracy.

At present, interpretability has no formal technical meaning. We define interpretability as the ability to explain or to present steps and results in understandable terms to a human. The demand for interpretability arises when there is a mismatch between the formal objectives of supervised learning (test set predictive performance) and the real world costs in a deployment setting. We will try to understand when this mitsmatch can arrive and how to measure it, when it is possible.

Following the literature we observe that some papers equate interpretability with understandability or intelligibility, others motivate interpretability by suggesting it to be a prerequisite for trust. Trust might also be defined subjectively. For example, a person might feel more at ease with a well understood model, even if this understanding served no obvious purpose. Trust might also denote confidence that the model will perform well with respect to the real objectives and scenarios. Another concept associated to interpretability relates to transparency, i.e., how does the model work? or what else can the model tell? Informally, transparency is the opposite of opacity or blackbox-ness and connotes some sense of understanding of the inner workings of the model.

Algorithms transparency and interpretability generate questions for the researchers due to the recent wish of enterprises to use deep methods provoking an explosion of the number of parameters to consider and consequently creating black boxes. Such methods need to be ”interpretable” to be used for important decisions. Choices and errors must be understood at least a posteriori (e.g. for court decisions, car driving, medical diagnoses or regulatory purposes). Although unfair treatments could be detected by analyzing the output of a model, a greater transparency could help preventing such issues and better control model bias a priori. In another hand, excessive transparency with outsiders can lead to business leaks, and a good balance is thus required to protect the secrets of manufacture of the companies. This however has to be distinguished with the transparency of the model for internal control, which only has a positive effect. Some of these points are described in the following.

1. Some models can be more interpretable, like sparse or low-dimensional linear models (regression, logistics, SVM), small decision Trees, Naive Bayes, or k-NN. The interpretability question concerns how to use more complex models. The complexity can arise from the size of the models, for instance in a regression Tree, the use of a big number of leaves will make the model hardly interpretable. Thus, in case of complex models, the traceability of the successive results can make them more interpretable. Also, visualizations of intermediate results such as the matrices of weights for convolutional networks can provide great insights.

2. Concerning algorithms developed on specific platforms[8] and the difficulty to make it in production in an enterprise, the need of transparency is crucial to avoid mis-understanding. The choice of the data indeed plays an important role but also the knowledge of the model behind the algorithm to be sure to obtain the expected accuracy. For instance, we can observe an overffiting in the results, due in part by the presence of spurious regressions. Indeed if the user has a limited knowledge in data analysis (statistics), the results could be spurious by a bad use of the algorithm or by a bad interpretation of the results[9]. The main point remains to determine the adapted person to be in charge of the interpretation of results. To produce a good explanation, the practicians must have a good understanding of the models[10]. The interpretation of the results also depends on the data used. Thus, the person in charge of verifying the quality of the data and the one that have a complete knowledge of the data base (if different) might be involved in the discussion of the results. Solutions such as Databricks aim at integrating the whole chain of data scientist, data analyst and data engineer within a common hub. These remarks open the question of the governance.

4.2 Governance and Human-machine interaction risk

Thus, whatever the choice of the use of big data and machine learning processes used in a enterprise to attain some objective, human will be a part of the process. Thus in this section we question the responsabilities associated to the decisions which are taken in fine: they concern the choice of the data and also the choice of the models and the decisions taken.

Currently, humans interacting with algorithms are not necessarily experts in the machine learning domain, thus its seems that a majority of models used in practice are solely relatively basic. Machines in the current economic state seem mostly used to augment humans' capabilities rather than for replacing them. In the future, the risks will come of an mis-understanding of the process depending of the persons in charge of the management of these algorithms and the persons designed to take decisions. Thus, a major recommandation concerning the use of algorithms in order to take decisions in an enterprise would be the obligation of the persons in charge of this process to understand how the algorithm works and how the results are provided.

- Auditing the processes is mandatory for companies and will be done also when machine learning and big data sets are used. It will be the occasion for the management of an enterprise to define an understandable process including the data set used, the algorithms used, their accuracy and explanation permitting to justify the taken decisions: implementation of back up chains (with data used, different steps of the algorithsm, estimated weights) could become the norm.

[8] PoC by providers

[9] to prevent overfitting from non-data-expert users, some platforms handles the data fitting in back-end, and are thus destinated to domain experts rather than data scientists

[10] for instance, to understand if the an algorithm is trained to detect correlations or causalities? Indeed, the explanation of the result of a model easily lead to suggest a causality effect between the variables

- Role of the regulation. The regulator can verify if a set of data discriminates. Verifying this equity property should be devoted to the regulator. Even if it can exist some relationship between equity and efficiency, this former point need to be discussed. Efficiency is linked to the objective function and measures through the accuracy of the algorithm, to obtain equity some constraints can be introduced in an algorithm.

- Although the algorithms are developed on external platforms the decisions still have to be made by the firm. It exist risks of reputation and governance as we have already alerted for data storage. These risks are generally not new because often firms use outside companies for specific tasks.

We point that, in fine, it is the human who takes decisions, no responsibility is supported by algorithms. The choice of the data is also determinant for the results. The regulator can help the enterprises in their choices, to provide efficient results based on interpretable algorithms and appropriate data sets. The compliance to the different laws on privacy and data properties need to be known by the users and intergrate in their learning processes. Regarding the property of the algorithms and their use, the law has not yet treated all the subjects and further evolutions are expected.

- As algorithms are considered "self-learning softwares" they are protected by the standard regulation on softwares. More complex issues are related to the property rights on patterns, since they are the result of both the algorithm (owned by the software company) and the data (owned by the database producer). This uncertainty could require the mentioning of rights on patterns in the contract established between the different parties.

- Insurances could also protect related events such as the reputation risk in case of accidents with the algorithm. However, who is responsible for unintended consequences? Is it the software company (entity), its director or the developer? Also, who is responsible in case of an accident following a hardware failure? Would the software company be blamed for failing to predict its own breakdown, or for lacking a security system that should handle these situations? Is it possible to measure this kind of risks? What could be a good framework? All the questions remain open and legislation in all countries need to be clearer on all these subjects.

4.3 Operational risks

The Basel II Committee defines operational risk as: "The risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. This definition includes legal risk, but excludes strategic and reputational risk." However, the Basel Committee recognises that operational risk is a term that has a variety of meanings and therefore, for internal purposes, banks are permitted to adopt their own definitions of operational risk, provided that the minimum elements in the Committees definition are included. With the arrival of the use of machine learning, new operational risks emerge that can be due to mistakes in the implementation, the choice and training of a model, the natural covariance shift and shocks on the data, or a systemic convergence of models. The algorithms can also be used to detect failures and anomalies in other systems like bugs, frauds, extreme and cyber-risks (including frauds involving debit and credit cards, unintended insiders, hacking or malware, documents, smartphone, hard drive, computers stolen, unintended disclosure among others). We now analyze more of these new specific events, and how to circumvent them.

4.3.1 Detecting algorithms failures

We propose some classifications of these risks without being exhaustive.

- Implementation. The detection of anomalies in the implementation of algorithms (bugs, etc.) is a hard task since computer methods evolve at a fast pace. It is indeed difficult to characterize a type of error that continuously changes. What unknown errors could happen, so far as unpredictable as black swans? How could an algorithm try to predict new kinds of errors by taking into account the changes applied to infrastructures? A meta learning detector trained on the distribution of the evolution of errors could help in their detection if their is a sort of stationarity in their evolution. Generally human errors are of a simpler nature, algorithms conversely introduce specific and complicated functions.

- Transition to production. The failure of an algorithm is relatively controlled during the training phase, as performances and errors are almost necessarily handled. However, depending on the type of task and availability of data, estimating the risks of failure of an algorithm in production can be challenging. Indeed, the training data might be insufficient to approximate the distribution of the whole phenomenon (few samples, clean training set), or the underlying object might be naturally shifting in covariance (e.g. market prices). The detection of failures could eventually be enhanced by a greater transparency of the models, or by providing more attention to the issue, such as using algorithms to predict the failures of other models: these models could potentially be better at detecting failures than the algorithm itself if they transferred knowledge accross the different tasks.

- Specific frauds directly attack the models.

  - With evasion Attacks, tiny changes in input lead to misclassifications. The attacker is trained using the same optimization problem than the actual classifier. Thus, it is possible to harden the task of the attacker by working on more constrained problems (less sensible to attacks). For example, images can be manipulated by slighly modifying pixels (an inperceptible way for humans), whereas for a simple regression with few parameters manipulating some points might either be visible or not greatly impact the results.
  - Poisoning Attacks add malicious inputs that are wrongly labeled. These attacks are designed to (almost) not impact the accuracy on the validation set, but eventually have large consequences on the test set accuracy. Although these attacks are incidious, they are relatively easy to detect since malicious samples must be different, by construction.
  - Privacy attacks consist in a large range of violations from the stealing of a model (model-extraction attacks) to model inversion used to steal the data that trained the attacked model (such as faces or fingerprints in biometric systems). Differential privacy is a way to counter privacy attacks by incorporating some noise in the data, and thus increase the complexity of reconstructing the original sensitive data.

- Systemic risk. At a global scale, a risk could be generated by the emergence of similar models and in favouring best practices, which could create an excessive homogeneity and pro-cyclicity, originating a systemic risk. Indeed, models are largely distributed and open-sourced and the adoption of the same state-of-the-art models in equivalent businesses seems to occur. Decisions made on the same phenomenon could even modify its behavior and potentially increase the systemic risk on related environments, be it economic or car traffic domains. This kind of situations starts to be problematic, due to the up-to-date attraction of these subjects in the bank and also by the lack of formation of some stakeholders. Such a development requires risk measures of extremely dependent events, for which copulas or dynamic networks-based tools could be useful.

4.3.2 Using algorithms to detect anomalies

Due to their capacities to find complex patterns, machine learning mehtods can also be efficient for the detection of anomalies.

- Anomalies such as hardware and software craches could potentially be detected upfront. Solutions are currently developed to detect abnormal data behavior that could indicate the origination of a failure. These systems typically extrapolates from a rich and coherent data such as: (i) performance indicators (that would breach an unusual levels), (ii) contextual data, (iii) softwares logs (latency, availability, bugs, crashes), (iv) business impact indicators, etc. The current of dockered and serverless applications could increase the risk of cascading breakdowns throughout processes, hence the need for advanced models able to detect such insidious failures. An ultimate objective could be to obtain recommandations on possible solutions to treat the issue, and even automaticly apply these maintenances. [11].

- Frauds can also be detected using the aforementioned methods (e.g. credit card frauds). The main advantage of machine learning models over simpler methods lies in their ability to continuously learn and adapt to new information collected on recent fraudulent cases, increasing their capability to detect other abnormal behaviors, and to systemize control procedures.

- Detection of cyber-risks: by their capacity to process massive amounts of data, algorithms could both enhance the implementation and detection of complex cyber-attacks. Indeed, some current securities attempt to evaluate whether the user is a robot or a human, and AI by construction tries to imitate the human behavior. Thus, AI algorithms continuously break these shields by improving at recognizing visual text and objects for instance. On the other side, improvements in AI also improve in the detection of fraudulent behaviors. The duality of the security problem is thus enriched by machine learning methods. Classical methods to measure these risks are based on extreme value distributions like for instance Generalized Pareto Distribution that can be reinforced by techniques combining networks and adversarial risk analysis for instance.

- Concerning extreme risks, models can be trained to understand the distribution of data, then extreme events can be measured, for instance with the VaR[12]. The algorithms could also be used to measure and coherently control the skewness and kurtosis, and thus generate stress tests by exagerating the normal conditions of the given environment. Hence, stress-GAN-tests[13] could be implemented to strain operating models on plausible extreme scenarios. This solution would follow the Basel III accords' philosophy, which introduced stress tests to verify the safety of banks in case of important shocks.

Some references related to these topics are now listed. An overview for machine learning model is provided in Kotsiantis et al. (2007). In Zhou et al. (2014) the complexity of the data associated to the their use with the algorithms is evocated, also in Chen and Lin (2014). The different facets of the GANs are developed in Goodfellow et al. (2014a). The theoretical underground of the models is discussed in Burrell (2016). Several interesting studies on model interpretability can be found in Lou et al. (2013), Kim (2015), Ribeiro et al. (2016c), Lipton (2016) and Doshi-Velez and Kim (2017), see also Oyallon et al. (2013) for an approach of the interpretability based on wavelet scattering transform, among many others. An interesting paper on regulating AI is Ebers (2019), even if it focuses also on robotics (subject that we do not investigated in this paper), see also Waltl and Vogl (2018).

[11] www.journaldunet.com/solutions/cloud-computing

[12] VaR: the Value at Risk is a common metric in the finance industry, used to measure the consequences of an extreme event for a given probability, i.e. by selecting a specific area of the distribution.

[13] GAN: Generative Adversarial Network are models that create plausible data, initially to cheaply increase the number of samples of a data set. There now exists a large panel of uses from attacks, image-video editing, models generalization and robustness, etc.

5 The notion of bias in machine learning

In the previous sections, the role of the data and the questions relative to the emergence of risks through the use of algorithms were analyzed. The results provided by machine learning systems are now investigated in relation with humans responsibilities: indeed these results are the keys for decision making. We begin to understand why they are very often qualified as "bisased" results.

It is important to analyze the sense given to this term and the context in which they are provided. Indeed, several meanings to "bias" can be distinguished: First bias can qualified statistical results. This property represents the inadequation of the data or model to reach a desired statistical objective. Second bias can be synonym of discrimination. For example, when a person does not reveive a loan, the question is : are the data set used to exclude or not this client for specific reason? In the recidivism prediction task, some procedures yield also at discriminating results. Third, cognitive biases is associated to social psychology. A cognitive bias is a mechanism of thought that causes a deviation of judgment. To take action or make sense of an event, the brain will use subconscious subjective beliefs. In certain domains this kind of bias can create specific behaviors like economic biases. Indeed, cognitive biases are the source of various anomalies affecting economic behavior and the efficiency of markets. The machines are indeed not directly subject to these biases, but indirectly via the choice of functions and factors by the humans building the model.

1. Statistical approach.

   A global process associated to AI is based on machine learning architectures. Such architectures include related-regression and classification procedures which can be based on supervised, unsupervised or reinforcement learning, using neural networks, Trees, SVM and k-NN models among others, as depicted in section 2.2. For supervised learning, a (complex) function f is defined such that given an input data set $X = (x_1, x_2, \ldots, x_n)$, an output vector $\hat{Y}$ is approximated: $\hat{Y} = f(X)$. From a statistical approach, this output vector $\hat{Y}$ is a random variable, and a bias can be measured by comparing this output to the true expected value $Y = (y_1, y_2, \ldots, y_n)$. A simple way to define this bias is to verify if, given (X, Y):

   $$E[\hat{Y}] = Y \tag{1}$$

   where the expectation is computed for the distribution which characterizes Y given X.

   The bias of the output $\hat{Y}$ can be measured given a specific target. In that case, "bias" is synonym to "error". Thus, the bias comes from an uncaptured part of the phenomenon. It can come from an inappropriate model f, from an insufficient information in X to predict Y, or from a mismatch between the information set (X, Y) relatively to the choice of the model f. Although a high bias is not desired, it has to be balanced with its variance conterpart, and often lead to a bias-variance thread off. We are going to analyze these three steps: the inadequate data set used for a specific objective function, the model risk and its computation and the bias detected on the output.

   (a) The bias due to the methodology or the algorithm is called a model selection problem. It can be solved using several adequation tests adapted to each algorithm for the specific objective. As a problem can often be solved by several models, a ranking of the algorithms can be derived to avoid the use of an algorithm that is not adapted to the task. The model finally selected might be the winner of this competition regarding metrics such as accuracy and adequacy. Thus, it is possible to reduce an error (or bias) originated by an algorithm

by comparing the results obtained with several other algorithms. In a first step we can say that the bias of an algorithm can be defined as a deviation between the true objective and the results as we define in (1).

(b) The bias due to the data can come from several sources: (i) data quality and representativeness; (ii) too small or unbalanced data set. Semi-supervised learning or SMOTE-like methods can help improving the generalization abilities, but for strongly imbalanced problems such as credit-scoring or recidivism that are often too specific for gathering data from similar problems, the predictability rates of these methods need to be more compelling.

2. Bias relative to discriminate Some applications are more prone to discrimination than others. For example, models used to inform judges about the likelihood that a defendant will re-offend if released use personal (although not necessarily identifying) data about individual defendants (like the race or the place where he/she lives). Given the importance of decisions regarding an individuals personal liberty, it is imperative that any input to the decision-making be fair with respect to legally or socially protected classes such as race, gender, sexual orientation, revenues etc. Unfortunately, some algorithms used to predict the recidivism risk of a specific population seem to have yielded discriminating results as it has been illustrated in several documents. We want to emphasize that, in fine, since the humans build the whole process, he/she remains responsable of the result and then of the decision making, for instance according to the Charter of Fundamental Rights of the European Union.[14] We can distinguish several schools of thinking.

   (a) One school of thought does not focus on a particular metric of fairness, but rather assumes a model will be fair if the protected (sensible) variable(s) are omitted from the analysis. But even if the protected variables are omitted, their effects will remain in the estimated model via their correlation with the permitted variables. In the case of regression models, this is known as omitted variable bias.

   (b) The second one defines fairness in terms of equivalence of some measure of predictive accuracy among all classes in a protected variable.

   (c) In the third one a model is typically considered fair if differences in the distribution of the models predictions conditional on the protected variable do not exceed some predetermined threshold, as measured by some appropriate notion of distance between probability distributions. In most cases, the allowable difference is zero, which is equivalent to the requirement that the predictive distribution is independent of the protected variable. This notion is sometimes called statistical parity or demographic parity.

   This discussion on the bias introduced by the analysis of the choice of the data cannot be dissociated to the analysis of the bias characteristic which is also given to the algorithms. Indeed, the notion of bias associated to an algorithm depends on several factors and is not related to only one contribution.

   The bias can be originated by several elements.

   (a) Data. The properties of the data can be at the origine of several bias. Indeed, the performance in production could be deteriorated, for instance, by an induced bias from (i) poor quality in source and format of the training data set, (ii) transformations done during the data pre-processing, (iii) mismatch with the format of the training set and, (iv) malicious transformations on the data.

[14] CFR: In France, articles 225-1, 225-2 et 225-4 of the criminal law define discrimination as a distinction that can operate on physical persons given their origins, their sex and their family situation.

(b) Weights. One key point for algorithms used in machine learning is the choice (or the estimation) of the parameter (or weights), thus algorithms can be manipulated insidiously by amending the value of their weights. Since deep models have thousands of weights, few modifications could be imperceptible but have an important impact on outputs produced. As it could modify performances, this can steer the result one way or another. It would potentially wreck the whole (costly) learning of an algorithm. Solution could be to implement means of control and restoration. In addition to secure connexions inside the algorithm, independent model could verify that weights are always amended coherently (the process needs to be explicit).

(c) Target objective. In practice, it is important to chose the data and model in adequation to the objective. For example, for the objective of recognizing objects on images, a convolutional neural network model can be trained on images containing the objects of interest. As a counter example, for the objective of predicting financial asset prices, a convolutional model trained on images would (i) not be able to use prices as inputs (since it would standardly require inputs of the same shape than during the training phase, for instance 28x28x3 squared colored pictures), (ii) if the model can still input prices, it would perform very poorly since patterns in images could hardly help in predicting prices. It is thus important that the users know the exact typology of the algorithms for their use case.

(d) Calibration. Poor calibration can lead to under-fit the data set impacting the performance in production.

Some of the questions we have analyzed here are also linked to the process which has to be developed to make an algorithm interpretable, following the discussion proposed in the previous subsection. Comparing the process done by an algorithm with the process which could be done by a human "by hand", the main differences lies in the computation speed and the capability of the algorithm to treat a large amount of data. But as the code of the algorithm is written by a human, and the information set X is also chosen by a human, even with an "unsupervised" algorithm, the human should be the responsible for the results generated. Indeed, even in that latter case, the paths to integrate new data have been, at the origin, oriented by a human. An algorithm cannot be qualified as discriminating, since the algorithm is not a human. Coherently, as the human built the machine, the human should be responsible of its acts, even indirectly executed by a machine, the human is thus indirectly acting discrinatively. Since algorithms are mainly used by companies, the responsibility could then be assumed by the social status of the company.

Further investigations of the problem of measurements and bias in complex environments are conducted in Tversky and Kahneman (1974), Angwin et al. (2016), and Kleinberg et al. (2016). It is specifically studied regarding discrimination when applied to recidivism predictions in Dieterich et al. (2016), Dressel and Farid (2018), Johndrow et al. (2019), and concerning other inequalities in ONeill (2016). In that context the notion of "fair" model is investigated in Clarke (2005) and Zafar et al. (2017), see also Kleinberg et al. (2016).

6 Discussions

As we illustrated previously, regulation and risks associated to big data and AI need to be analyzed at different levels. Due to the study developed in this paper, it is not clear whether a new regulation is necessary and if it would be adapted now or even later to the challenges created by the use of big data and AI. There is a feeling of a lack of theoretical understanding in the research community, although some deep methods are already efficient in production, hence the difficulty to propose specific strategies in terms of risk measurement. As it is important not to obfuscate these risks and issues, we

introduce some discussions and paths toward solutions addressed by researchers and practitioners. We also emphasize the necessity of simultaneous philosophical, ethical and political debates alongside the question of risk measurement.

6.1 Focus on data sets

In the following we discuss some of the most intriguing risks related to (i) the data representativity, and (ii) regulations around data property.

- Data representativity. We first emphasize that different communauties have distinct appreciations of the problem: developers and researchers have been looking for features in the data that is representative enough make accurate predictions, whereas lawyers have been more concerned with the understanding of the internal logic that could assess their lawfulness. Nevertheless, "distorsions" observed in the result that can be qualified as "bias" or "discrimination" correspond to the same problem. Several approaches can be used to assess the "representativity" of a data set for a particular objective.

  1. *Statistical samples* can be used to control a facet of the representativity. Working with big data sets does not prevent from applying these methods and using any available data might rarely be sufficient to avoid "bias" in the data. Indeed, volume only reduces the subset bias, and does not correct any discriminatory pattern intrinsic to the data generating process.
  2. *Statistical parity* can solve the problem of protected data (as race, gender, or other variable). The approach creates an adjusted set of covariates that are independent of a (set of) protected characteristic Z. More formally, it creates maps that optimally transform univariate X_j of the information set X to $\tilde{X}_j$, such that all information about Z (a potentially multivariate set of protected variables) is removed from $\tilde{X}_j$. By applying this method to each X_j independently, provides a way to achieve pairwise independence between each X_j and Z. Suppose we have a response $\hat{Y}$ and predictors (Z, X), where Z represent protected characteristics. Let be the following model:

     $$f : X \to \hat{Y}. \tag{2}$$

     The goal is not to use any information about Z in predicting $\hat{Y}$: this permits to define a fair prediction rule with respect to the protected characteristics Z if and only if $\hat{Y} \perp Z$[15]. Although f is not a function of Z in (2), this is insufficient to guarante $\hat{Y} \perp Z$ unless $X \perp Z$. In the overwhelming majority of applications, X and Z are dependent, and thus we must take additional measures to ensure $\hat{Y}$ is fair. To achieve this point, we proceed using an algorithm for transforming a univariate X to $\tilde{X}$ such that $\tilde{X} \perp Z$ with minimal information loss. Thus if the transformation is correctly done, the result will not be as bias as without transformation. Thus if the algorithm used on this data is adequate the result will be unbiased and non discriminating or "neutral" in the sense of the objective chosen function.
  3. *Equalized odds* are another approach consisting in verifying the existence of a parity in two groups applying the following rule. Using the same notation as before, $Z \in \{0, 1\}$ being still the discriminating variable and $\hat{Y} \in \{0, 1\}$ the predicted variable. We seek to learn to predict outcomes that are accurate with respect to Y but fair with respect to Z. One way is to ensure that the positive outcome is given to the two groups at the same rate, i.e. $P[\hat{Y} = 1|Z = 0] = P[\hat{Y} = 1|Z = 1]$. However, the usefulness of this constraint can

[15] $\perp$ means independence

be limited if the base rates of the two groups differ, i.e. if $P[\hat{Y} = 1|Z = 0] \neq P[\hat{Y} = 1|Z = 1]$. In this case, we can pose an alternate criterion by conditioning the metric on the ground truth Y, yielding equalized odds and equal opportunity.

4. Unbalanced data sets. The performance of machine learning algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the costs of different errors vary markedly. The machine learning community has addressed the issue of class imbalance in two ways. One is to assign distinct costs to training examples. The other is to re-sample the original dataset, either by over- sampling the minority class and/or under-sampling the majority class. The Smote algorithm creates an over-sampling approach in which the minority class is over-sampled by creating synthetic examples rather than by over-sampling with replacement. In this algorithm the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors we consider. For instance, depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen.[16]

5. Counter factual explanations differ from the previous approaches, which analyze the intrinsic properties of data sets. It consists in making changes on some variables and measuring the impact on the output $\hat{Y}$. The counterfactual approach permits to provide evidence that an algorithmic decision is affected by a protected variable, and that it may therefore be discriminatory. Counterfactual explanations are crafted in such a way as to provide a minimal amount of information capable of altering a decision, and they do not require the data subject to understand any of the internal logic of a model in order to make use of it. With the counterfactual approach, evidences can be provided on the (direct or indirect) use of a protected variable by the algorithm, which may therefore be discriminatory. For some types of distance function, if the counterfactuals found change some variable Z, then the treatment of that individual is dependent on this variable. Counterfactuals describe only some of the dependencies between a particular decision and specific external facts. The underlying idea is quite simple: the idea is to find a neighbor from the input which provides a different prediction with the same classifier. *It could provide some way to re-thinking some results and be an help for the regulator to avoid bad results.*

   The previous notations are used, with $X = (x_i)$ denoting the input, $Y = (y_i)$ the output, and representing standard classifiers of machine learning f_w trained by finding the optimal set of weights w_i that minimises an objective loss function $l(.)$ over a set of input data X, then the objective is to compute:

   $$argmin_w l(f_w(x_i), y_i) + \rho(w)$$

   where ρ is a regularizer over the weights. The idea is to find a counterfactual x' as close to the original point x_i as possible such that $f_w(x')$ is equal to a new target y'. We can find x' by holding w fixed and minimizing the related objective:

   $$argmin_{x'} max_\lambda \lambda((f_w(x') - y')^2 + d(x_i, x'),$$

   where $d(,)$ is a distance function that measures how far the counterfactual x' and the original data point x_i are from one another. In practice, maximisation over λ is done

[16] The implementation of the Smote algorithm currently uses five (k=5) nearest neighbors by default. Synthetic samples are generated in the following way: take the difference between the feature vector (sample) under consideration and its nearest neighbor; multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration; this causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general.

by iteratively solving for x' and increasing λ until a sufficiently close solution is found. The choice of optimiser for these problems is relatively unimportant, in counter-party the choice of the distance d is important. Depending on the data it could be the L^1 or L^2 norm, the Manhattan distance weighted by the inverse median absolute deviation.

As counterfactuals provide a method to explain some of the rationale of an automated decision while avoiding the major pitfalls of interpretability or opening the black box, *it may prove a highly useful mechanism to meet the explicit requirements and background aims of the GDPR*. This can concern subjects like racism, profiling, gender, etc. In that sense counterfactuals represent an easy first step that balances transparency, explainability, and accountability with other interests such as minimising the regulatory burden on business interest or preserving the privacy of others, while potentially increasing public acceptance of automatic decisions.

6. The regulator can verify, depending on the objective of an algorithm, the existence of bias originated by the data set by investigating (i) correlations with a protected variable, (ii) unexpected unbalanced subsets, (iii) correlation between variables, etc. and apply some of the previous methodologies to identify and prevent these biases to impact the result. Complementary procedures coud enhance the understanding of a data set by taking into account their meta data (distribution of related factors, extraction method, etc.). It is expected, during the auditing process, that the data should be provided and that an extensive inspection of the quality of the data set, and their good "representativity" could be conducted using classic statistical tools including AFC, clustering, or metrics tailored for specific types of data.

- Legal risk. For companies, the regulation related to big data security and privacy has already been discussed, and in some cases solutions do exist. In Europe, the General Data Protection Regulation (GDPR) was enforced on 25 May 2018, and organizations that are not compliant could now *face heavy fines*. The website https://eugdpr.org/ helps enterprise to become GDPR compliant. However, regulators might have trouble controlling deviances as a gap exist between current legal tools and the cutting-edge technologies that are shaping our societies and ourselves. If the European website is a *resource to educate organizations* about the main elements, it appears indispensable *to cultivate ethical sensibilities around information technologies*. Experts must be trained on ethics, privacy, right to be forgotten, right to data expiry, ownership of social graphs. Although there is a noticeable increase in attention on these topics, it is also the role of the politics and professionals of education and formation to integrate those notions in their syllabus formations.

  Applicable law is not the same for each country and region of the world. There is no concensus on the treatment of data and it is unclear to which regulatory framework one should refer in many cases. For instance in China, the governmental approach seems different to European rights concerning privacy. The use of private data in China is explained in particular to "increase access to finance of low-income families and micro-enterprises in China" and for Chinese government it is natural to use citizens private data. In the United States, the protection of American residents' data is regulated on both the national level through the Cloud Act[17] and the state level. In Europe, the paragraph 71 of the recitals (the preamble to the GDPR) explicitly

[17]The Clarifying Lawful Overseas Use of Data Act or CLOUD Act (H.R. 4943) is a United States federal law enacted in 2018. The CLOUD Act amends the Stored Communications Act (SCA) of 1986 to allow federal law enforcement to compel U.S.-based technology companies via warrant or subpoena to provide requested data stored on servers regardless of whether the data are stored in the U.S. or on foreign soil.

requires data controllers to implement appropriate technical and organizational measures that prevents, inter alia, discriminatory effects on the basis of processing sensitive data[18]. Thus, in Europe the role of the regulators is concerned by this law and the way to apply it. Therefore, wherever multiple legislations are applicable about the privacy of data, political discussions need to be engaged.

6.2 Focus on algorithms

Interpretability and adversarial attacks on algorithms are important matters for both companies and regulators. Unexpectedly, adversarial perturbations can serve both issues. In the following, methods to evaluate algorithms and interpret their results are first studied, then techniques to counter attacks are investigated.

- Interpretability is useful for compliance (right to explanation), but also for privacy, fairness, robustness and trust. One can say that "explanation is closely related to the concept of interpretability: systems are interpretable if their operations can be understood by humans, either through introspection or through a produced explanation. In machine learning, explanation is often a difficult task since most models are not readily interpretable"[19]. Interpreting a model predictions requires the understanding of the whole process: (i) the model selection, (ii) the training, (iii) and the evaluation of the result. The quality of a model can only be assessed relatively to a particular metric, which is expected to summarize the interests of the organization using the algorithm. Thus, the business success criterion needs to be converted into a predictive modeling criterion so the modeler can use it for selecting the models. If the purpose of the model is only to provide highly accurate predictions, accuracy measures might be sufficient. However, if interpretation is what matters the most for the business, metrics assessing the interpretability of the model might be used jointly. Thus, slighly less accurate but more transparent models would be selected. In the following, classical model evaluation metrics are first analyzed and explainability methods are then investigated.

    1. The correct evaluation of an algorithm for a specific objective is crucial. Since multiple modellings could reach the same accuracy while performing differently according to other metrics, the choice of the measure is thus of critical importance. A large panel of criteria can be used, from which: the Percent Correct Classification (PCC: a technique for combining the results of structural classifiers), confusion matrices providing a summary of different kinds of errors (Type I and Type II errors, precision and recall, false alarms and false dismissals), specificity and sensitivity, etc.

       PCC and confusion matrices are appropriate measurements when an entire population must be scored and acted upon. If one treats a population subset, sorting the population by model score and acting on only a portion of those entities in the selected group can be

[18] GDPR Paragraph 71: "In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organizational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimized, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject, and prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect. Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions."

[19] Biran and Cotton (2017)

accomplished through metrics such as Lift, Gain, ROC, and Area Under the Curve (AUC). For continuous-valued estimation problems, metrics often used for assessing models are R^2, average error, Mean Squared Error (MSE), median error, average absolute error, and median absolute error. Average errors can be useful in determining whether the models are biased toward positive or negative errors, whereas Average absolute errors are used to estimate their magnitude (be it positive or negative).

Analysts most often examine the entire range of predicted values by considering scatter plots of actual versus predicted values or actual versus residuals (errors). Another candidate for customized scoring of models includes Return On Investment (ROI) or profit, where there is a fixed or variable cost associated with the treatment of a customer or transaction (a record in the data), and a fixed or variable return or benefit if the customer responds favorably.

In summary, the task of assessing models requires an in-depth understanding of the metrics. Analysts need to understand the intent of the models and match the metrics accordingly. The investigation could use global metrics that treat every record alike, such as R^2 or PCC, or rank-ordered metrics such as lift at a predetermined depth or an ROI calculation. The metric used to assess the performance of a model can have more influence on the result than the algorithm itself. The regulator will be in charge of deciding the required level of analysis relative to the accuracy of models used.

2. Interpretability by local methods. Evaluation of a model have been discussed, but it also exists a trade-off between the performance of the model and the effort required to interpret it - especially in complex domains like text and image analysis, where the input space is very large. In these contexts, accuracy is usually sacrificed for models that are enough compact and transparent to be comprehensible by humans. Some models are simple, intuitive and are thus easy to understand. For complex models such as ensemble methods or deep networks, the base model requires an additional module to generate an explanation, which can yield interpretable approximations of the original model.

   Using the same notations as before, f is the original prediction model to be explained and g the explanation model. Here, we focus on local methods designed to explain a prediction $f(x)$ based on a single input x. We consider a simplified inputs x' (as in counterfactual explanations) that map to the original inputs through a mapping function h: $x = h_x(x')$. We want that $g(z') \approx f(h_x(z'))$ whenever $z' \approx x'$. (Note that $h_x(x') = x$ even though x' may contain less information than x because h_x is specific to the current input x.) In case of binary variables, an explanation model g is defined as:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

   where $z' \in \{0,1\}^M$, M is the number of simplified input features, ϕ_i are parameters and belong to $\mathbb{R}$. Methods with explanation models g attribute an effect ϕ_i to each feature, and summing the effects of all feature attributions approximates the output $f(x)$ of the original model. Two methods based on this approach are proposed below.

   (a) Related LIME methods. The LIME method interprets individual model predictions based on locally approximating the model around a given prediction. LIME refers to simplified inputs x' as interpretable inputs, and the mapping $x = h_x(x')$ converts a binary vector of interpretable inputs into the original input space. Different types

of h_x mappings are used for different input spaces. For bag of words text features, h_x converts a vector of 1s or 0s (present or not) into the original word count if the simplified input is one, or zero if the simplified input is zero. For images, h_x treats the image as a set of super pixels; it then maps 1 to leaving the super pixel as its original value and 0 to replacing the super pixel with an average of neighboring pixels (this is meant to represent being missing). The important point concerning this approach is that it is based on local accuracy meaning that when approximating the original model f for a specific input x, local accuracy requires the explanation model to at least match the output of f for the simplified input x' (which corresponds to the original input x) such that:

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i.$$

To find ϕ, LIME minimizes the following objective function:

$$\xi = argmin_{g \in G} l(f, g, \pi'_x) + \phi(g)$$

where l is a loss function, π'_x a kernel and ϕ a function which penalizes the complexity of g. Several approaches have been developed in choosing the parameters in the previous equation and depending how to find the loss function l, weighting kernel π'_x, and the regularization term, for instance the deepLiFT method (adapted for deep learning models for which the mapping h_x is not rectricted to binary values in its converting) or Shapley regression with the SHAP algorithm values (used in presence of multicollinearity inside the inputs features).

(b) Model-agnostic approach. The model-agnostic explanation procedure provides explanations that are local in their scope, and could be a way to solve the question of interpretability. It permits to use more powerful models, providing better explanations, with flexibility and coherent explanations. This approach is based on rules intuitive to humans, requiring low effort to comprehend and apply, explaining individual predictions with if-then rules. The idea is the following: considering the inputs (X, Z), the output Y, and the model explained through a function f, one introduces some constraints on the variable X, and defined C, the set of constraints $c \in C_x$ (c could be education for instance). One assumes that one can sample inputs Z where the constraints C are met, and gets a distribution of interest D sampled from $D(z|c, x)$. Then the objective function becomes:

$$Precision(f, x, c, D) = E_{D(z|c,x)}[1_{f(x)=f(z)}].$$

In order to balance precision, coverage and effort, the function defined previously is optimized finding the minimum constraint c to get the best result, such that:

$$min_{c \in C}|c|s.t.Precision(f, x, c, D) > 1 - \epsilon.$$

- In adversarial approaches, algorithms capable of computing counterfactuals are used to confuse existing classifiers by generating a synthetic data point close to an existing one such that the new synthetic data point is classified differently than the original one. This approach has been mainly developed for computer vision, but also for data sets for which fairness is important for instance concerning demographic parity, equality of odds and calibration. For these latter cases, applications can be in recidivism, creation of profiling or scores on individuals. We can

distinguish (i) adversarial example which is a modified version of a clean set that is intentionally perturbed (e.g. by adding noise) to fool a machine learning model, such as deep convolutional neural networks, (ii) adversarial perturbations can be the noise that is added to the clean set to make it an adversarial example, (iii) adversarial training using adversarial set besides the clean set to train machine learning models.

1. Demographic parity (e.g. discrimination). In this approach, the idea is to add an adversary to a network that predicts demographic parity in order to counteract racial biases found in criminal history datasets for instance.

   The model structure is the same as before, with f denoting the multi-layer neural network that provides an output $\hat{Y}$, Y being the ground truth. The idea is to obtain the output $\hat{Y}$ satisfying demographic parity or equality of odds for the demographic variable Z. Even if Z is not an input feature to the neural network, it may be correlated with other features, from which the network can learn a bias. For instance, we input the logit model from f to an adversarial neural network model g that learns to classify demographic Z. If $\hat{Y}$ is biased for demographic Z, g should learn to have a high accuracy because the logit will be highly predictive of Z. Our goal is for neural network f to predict $\hat{Y}$ accurately and for g to predict Z poorly. If this is achieved, the model will be outputting an unbiased, accurate $\hat{Y}$. To reach this objective one can use binary cross-entropy losses for f and g, which we refer to as L_f and L_g respectively. To train g, L_f is back-propagated through g. At the same time, f is trained to be good at predicting Y and bad at predicting a logit that is highly correlated with Z. If we subtract L_g from L_f, f will be encouraged to maximize L_g, which will produce a logit that cannot be used to predict Z and $\hat{Y}$ values that are closer to achieving parity. This indicates that the adversarial model depends on more holistic information while the regular "bias" model is mostly dependent on charge degree, age, and priors.

2. Attacks. Some malicious algorithms generate samples (typically images and audio, but which can work on any input) that clearly appear to belong to a given class of object (for a human's eye), but is however misclassified by the attacked model. To train, models slightly modify input variables of a clean sample (e.g. pixels for images) and monitor the effects of the change to obtain the desired mis-classification. It exists a recent and prospective literature of adversarial attacks on Recurrent Neural Networks (RNN) on which attacks were originally conducted, but successfull tests also exist against state-of-the-art deep reinforcement learning algorithms (RL) and convolutional neural networks (CNN) on object detection and facial recognition problems or in physics.

   Currently, the defenses against adversarial attacks are being developed along several directions, for instance using modified training samples during the learning phase or modified input while testing the model, modifying networks by adding more layers or changing loss and activation functions, using external models as network adds-on when classifying unseen examples, or limiting the number of queries that can be asked to the model in production thus limiting reverse engineering.

3. Causality. The analysis of causality could be a fundamental component that misses to current algorithms, and which could be addressed by adversarial models. Indeed, the typical machine learning model observes data and finds correlations that hopefully hold in out-of-sample data. In that sense, the supervised learning approach is easily mis-interpreted as it insinuates that the link between inputs and the ouput (labels) is a causality. There-

fore, two components might be missing to converge toward a reflexion closer to humans', which are called sometimes "intervention" and "counterfactuals". The former introduces causality in the model, and corresponds to observing the effect of actions or "deliberate alterations", taking a step further from standard observations. The latter consists in making hypotheses on the causality in the data, which could intentionnaly be left aside to build an imaginary world that strives to assess what could have happened without the effect of the deliberate alteration. This last step builds a world that is not revealed in the data, and as new hypotheses based on these new assumptions are confirmed by new data, the model tends to understand the inner workings of the studied phenomenon. Put another way, it is similar to estimating the distribution of the data and augmenting the samples coherently using adversarial models. This analysis could be done using conditional probabilities that respect some axioms, defining how to translate a causal query into an equation.

Concerning all these discussions some references are now provided. Concerning the forbidden data, some specific study is provided in Hardt et al. (2016) for the method of equalized odds, and in Wachter et al. (2017) for the counterfactual approach. Model performances concerning for instance imbalanced data sets or accuracy criteria are discussed in Lewis and Catlett (1994), Pazzani et al. (1994), Kubat et al. (1997), Ling and Li (1998), Japkowicz et al. (2000), and Chawla et al. (2002), for a general overview we refer to Saporta (2006). Algorithms interpretability's solutions are investigated in Lundberg and Lee (2017) and Ribeiro et al. (2016b) for the model-agnostic approach, and in Ribeiro et al. (2016a) and Wadsworth et al. (2018) for the adversarial method. Causality issues can find an interesting study in Pearl and Mackenzie (2018). Attacks on models are developed in many papers, we cite for instance Akhtar and Mian (2018), Rozsa et al. (2016), Moosavi-Dezfooli et al. (2016), Papernot et al. (2017), Liu et al. (2011), Huang et al. (2011), Xie et al. (2010), Kurakin et al. (2016), and Goodfellow et al. (2014b).

7 Conclusion

In this paper was established a list of the main risks associated to the use of big data frameworks through models of artificial intelligence. Then, some of the recurrent difficulties experienced by companies, institutions, banks and insurance companies were analyzed more thoroughly. Discussions were also partly related to the recent interests that emerged from data privacy and ethical issues regarding regulation in several countries. The proposals can be seen as routes toward a response to the risks leveraged in the public consultation done by ACPR in 2019 - Fliche and Su (2018), and in the meeting organized in Paris jointly with ACPR, AMF and the European Fintech Project H2020, March 2019[20].

Some proposals are related to the treatments of data sets in order to avoid specific "biases" associated to quality, legal risks and forbidden variables. Concerning algorithms, mechanisms to ensure responsibility and accountability of their outcomes, associated to governance and decision risks are described. The need of a parallel investigation of risks along with their potential solutions is not new, with regard to best practices for decision making in companies, compliance and auditing vis-a-vis the regulation. The list of risks is not exhaustive and focused on the most discussed subjects. Technical research will continue due to the infatuation of the matter, as the regulation needs to be enlarged, and as political discussions are required at different levels to uniformize the uses within countries. The current regulatory framework is indeed not sufficient and needs to be more flexible, rigorous, and shared across countries.

[20] https://www.fintech-ho2020.eu/

Reflexions also point to a need for humans, and especially domain experts and statisticians throughout the process since fragmented frameworks, where the data extraction and model implementation are made by different companies, increase the risks of models misspecification, opacity and bias in the result. Some recent frameworks are working in that direction through the implementation of hubs[21], and we expect more global integrated frameworks to be developed.

We have not proposed a specific risk measure to take into account the risks listed and analyzed all along this paper. Indeed, the risks for which such risk measure can be associated such as operational risks (including cyber risks) are not new and documented in Bale III guidelines, even though the source of some data could have evolved. The way to measure risks remains based on standard statistical tools, largely documented in the literature. Most model risks are not new either, and can be assessed using existing metrics, excepted for specific risks leveraged by the opacity and randomness of the models, and we present some strategies involved to avoid them and measure them. For legal risks, which are prominent on this subject, sanctions do exist, have already been applied, and depend on the legislation of each country.
This paper investigating the construction and maintenance of infrastructures based on big data, their use and that of machine learning models open the debate around the questions of their risks highlighting the necessity to work with an holistic approach and recommends not to work in a piecemeal way. Although the challenges seem daunting, discussions between regulators, researchers and practicians would permit to better control these risks.

References

Addo, P., Guégan, D., and Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, 6(2):38. doi:10.3390.

Akhtar, N. and Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430.

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., and Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, 2(3):87–93.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *Propublica*. Link.

Banasik, J. and Crook, J. (2005). Credit scoring, augmentation and lean models. *Journal of the Operational Research Society*, 56(9):1072–1081.

Barry, L. (2019). Justice ou justesse ? l'équité de l'assurance. WP Chaire Pari.

Basel-I (2004). International convergence of capital measurement and capital standards: a revised framework. *Bank for international settlements*.

Basel-II (2016). Standardised measurement approach for operational risks: a revised framework. *Bank for international settlements*.

Basel-III (2017). Finalizing post-crisis reforms. *Bank for international settlements*.

Biran, O. and Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8.

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679.

[21] e.g. Databricks

Burrell, J. (2016). How the machine thinks: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512.

Calvano, E., Calzolari, G., Denicolò, V., and Pastorello, S. (2018). Artificial intelligence, algorithmic pricing and collusion.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, X.-W. and Lin, X. (2014). Big data deep learning: challenges and perspectives. *IEEE access*, 2:514–525.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict management and peace science*, 22(4):341–352.

Demchenko, Y., De Laat, C., and Membrey, P. (2014). Defining architecture components of the big data ecosystem. In *2014 International Conference on Collaboration Technologies and Systems (CTS)*, pages 104–112. IEEE.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*.

Ding, X., Zhang, Y., Liu, T., and Duan, J. (2015). Deep learning for event-driven stock prediction. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

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

Ebers, M. (2019). Regulating ai and robotics: Ethical and legal challenges.

Feng, G., He, J., and Polson, N. G. (2018). Deep learning for predicting asset returns. *arXiv preprint arXiv:1804.09314*.

Fliche, O. and Su, Y. (2018). Artificial intelligence: challenges for the financial sector. ACPR - Banque de France.

Galton, F. et al. (1909). *Essays in eugenics.[Part 1]*. The Eugenics Education Society.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014b). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Goodman, B. and Flaxman, S. (2016). European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*.

Graunt, J. (1662). *1662. Natural and political observations*. Royal Sociatey of London, fifth edition.

G'Sell, F. (2018). L'automatisation des décisions de justice, jusqu'où ? *Annales des Mines, Enjeux numériques*, (3).

Guégan, D. and Hassani, B. (2019). *Risk Measurement: From Quantitative Measures to Management Decisions*. Springer.

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

Huang, X., Gao, J., Buldyrev, S. V., Havlin, S., and Stanley, H. E. (2011). Robustness of interdependent networks under targeted attack. *Physical Review E*, 83(6):065101.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Japkowicz, N. et al. (2000). Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68, pages 10–15. Menlo Park, CA.

Johndrow, J. E., Lum, K., et al. (2019). An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220.

Jouini, E., Meddeb, M., and Touzi, N. (2004). Vector-valued coherent risk measures. *Finance and stochastics*, 8(4):531–552.

Kim, B. (2015). *Interactive and interpretable machine learning models for human machine collaboration*. PhD thesis, Massachusetts Institute of Technology.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography*, 3(3):262–267.

Klein, T. (2019). Autonomous algorithmic collusion: Q-learning under sequential pricing. *Amsterdam Law School Research Paper*, (2018-15):2018–05.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24.

Kubat, M., Matwin, S., et al. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, pages 179–186. Nashville, USA.

Kurakin, A., Goodfellow, I., and Bengio, S. (2016). Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.

Landau, S. (2015). Control use of data to protect privacy. *Science*, 347(6221):504–506.

Lewis, D. D. and Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

Li, Z., Tian, Y., Li, K., Zhou, F., and Yang, W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, 74:105–114.

Ling, C. X. and Li, C. (1998). Data mining for direct marketing: Problems and solutions. In *Kdd*, volume 98, pages 73–79.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Liu, Y., Ning, P., and Reiter, M. K. (2011). False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)*, 14(1):13.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631. ACM.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.

Nevmyvaka, Y., Feng, Y., and Kearns, M. (2006). Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd international conference on Machine learning*, pages 673–680. ACM.

Nunan, D. and Di Domenico, M. (2013). Market research and the ethics of big data. *International Journal of Market Research*, 55(4):505–520.

OECD (2017). Algorithms and collusion: Competition policy in the digital age. www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm.

Oyallon, E., Mallat, S., and Sifre, L. (2013). Generic deep networks with wavelet scattering. *Ecole Normale Supŕieure, Paris, France*.

ONeill, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. *Nueva York, NY: Crown Publishing Group*.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519. ACM.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., and Brunk, C. (1994). Reducing misclassification costs. In *Machine Learning Proceedings 1994*, pages 217–225. Elsevier.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.

Quetelet, A. (1846). *Lettres à SAR le Duc Régnant de Saxe-Cobourg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politiques*. Hayez.

Ram, J., Zhang, C., and Koronios, A. (2016). The implications of big data analytics on business intelligence: A qualitative study in china. *Procedia Computer Science*, 87:221–226.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the us stock market. *Journal of Banking & Finance*, 84:25–40.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). Nothing else matters: model-agnostic explanations by identifying prediction invariance. *arXiv preprint arXiv:1611.05817*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016c). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

Richards, N. M. and King, J. H. (2014). Big data ethics. *Wake Forest L. Rev.*, 49:393.

Rosenblatt, F. (1960). Perceptron simulation experiments. *Proceedings of the IRE*, 48(3):301–309.

Rozsa, A., Gunther, M., and Boult, T. E. (2016). Towards robust deep neural networks with bang. *arXiv preprint arXiv:1612.00138*.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, 63(2):129.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(October):433–60.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2):2018.

Wadsworth, C., Vera, F., and Piech, C. (2018). Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*.

Waltl, B. and Vogl, R. (2018). Explainable artificial intelligencethe new frontier in legal informatics. *Jusletter IT*, 4:1–10.

Xie, L., Mo, Y., and Sinopoli, B. (2010). False data injection attacks in electricity markets. In *2010 First IEEE International Conference on Smart Grid Communications*, pages 226–231. IEEE.

Xing, E. P., Ho, Q., Dai, W., Kim, J. K., Wei, J., Lee, S., Zheng, X., Xie, P., Kumar, A., and Yu, Y. (2015). Petuum: A new platform for distributed machine learning on big data. *IEEE Transactions on Big Data*, 1(2):49–67.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee.

Zhou, Z.-H., Chawla, N. V., Jin, Y., and Williams, G. J. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives. *IEEE Computational Intelligence Magazine*, 9(4):62–74.

Zwitter, A. (2014). Big data ethics. *Big Data & Society*, 1(2):2053951714559253.