



rev.research  
& advanced ML



# Regulating Machine Learning: where do we stand?

State of the Art and Challenges Ahead

# **1. Foreword**

AXA embraces Artificial Intelligence (AI) and Machine Learning (ML) in its business strategy because they provide powerful instruments to create value. This is a worldwide trend, across all industries, more and more processes are managed by these new technologies.

Recent advances in ML promise great opportunities. But the emergence of new technologies challenges existing regulation. In the EU, the use of ML is indirectly regulated by the General Data Protection Regulation (GDPR). It addresses the topic by strengthening data protection and privacy for all individuals; data heavily leveraged by ML. Since May 2018, data controllers, like AXA, are required to put appropriate technical and organisational measures in place to implement the new data protection principles.

With this document, we want to encourage all stakeholders to understand the most fundamental risks introduced by usage of ML and think ahead anticipating future changes in regulation. More specifically, we seek to explain to CDOs, DPOs, data scientists, actuaries, and any other interested parties how ML algorithms are different from conventional algorithms, in particular with respect to its relationship to data. We identify the most intrinsic risks when using ML at scale: fairness, reliability, explainability.

For each of these challenges we describe how current regulation is addressing them, followed by a discussion intended to put light on what is required by regulation and limitations of the current approach.

Moreover, although this document does not provide any mitigation tools per se, we have tried to expose the drawbacks, so that when deciding to put in production such a technology the right balance risk vs. benefit is considered for the best of interests.

A handwritten signature in black ink, appearing to read "Marcin Detyniecki".

**Marcin DETYNIECKI**

Chief Data Scientist



## 2. Table of content

1	Foreword	2
2	Table of content	5
3	Regulating Machine Learning: why should we care?	7
3.1	Machine Learning: a different type of algorithms	9
3.2	Machine Learning and personal data: introducing GDPR	13
4	What are the intrinsic challenges of Machine Learning?	15
4.1	Fairness: learning and reproducing bias	17
4.2	Reliability	23
4.3	Explainability	27
5	Concluding remarks	31
6	References	34



### **3. Regulating Machine Learning: why should we care?**

In recent years, Artificial Intelligence (AI) entered a new era. Enabled by an innovative type of algorithms called Machine Learning (ML) algorithms, by the multiplication of data sets and the tenfold increase in processing power, a wide range of applications has emerged, among others automated translation, autonomous car, cancer detection. This gives legitimate rise to hopes for the benefits this new technology will bring to our society.

In this white paper, we identify what we believe are the most fundamental challenges introduced by ML in its intrinsic nature. Other risks, such as misuse of ML, or malfunction resulting from inadequate or unfair input to these algorithms, will not be addressed. Even if those are critical issues, their nature is independent of ML and they are known for a long time. Thus, those cases are already well regulated. For instance, using ML for criminal or intentionally discriminatory purposes, is today covered by criminal or penal law.

As with any new development, besides the great potential of AI, there are also drawbacks associated – some of them yet unexplored. In order to ensure a sustainable success of the AI revolution, it is particularly important to at least roughly understand those risks.

Currently, AXA is fully compliant with the rules set by regulation. Still, we think this is no reason to rest. With this document, we would like to raise awareness of the particular characteristics of ML and stimulate forward-looking actions. Doing so will make our company ready for possible legal changes in the future, and

also enable us to continuously follow the ambitious values we have committed ourselves to.

As we will see in the following section, ML algorithms are strongly entangled with data, not only because ML needs data to execute, but also because ML is built, we may even say “grown”, from data.

In most situations, personal data will be used to train the ML algorithm. These data are subject of a special protection, at European level, mainly by the General Data Protection Regulation. The purpose of this regulation, which entered into force on 25th May 2018, is to harmonize at European level the conditions for the processing of personal data and their use, particularly for decision-making.

In the following, we start by providing some definitions and clarifications around Machine Learning contextual information such as more details on the General Data Protection Regulation. Then, we address three different challenges: fairness and bias, reliability and transparency, and explainability. For each challenge we provide a simplified explanation, followed by the responses provided by regulation today, mainly GDPR. Based on the current state, we then open the discussion by presenting what we see as limitations of these answers. Our objective is to raise awareness and initiate a fruitful exchange of thoughts in order to anticipate any potential risks well ahead whilst still benefiting from AI.

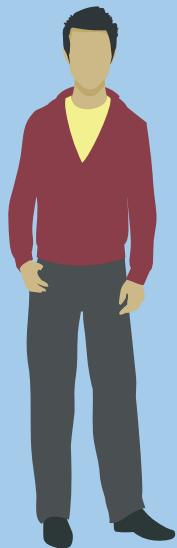
The reader could expect to find recommendations on how to mitigate the above-mentioned risks. We hope that, after reading this document, he or she will fully understand why for each of the challenges presented there is no obvious and simple solution.

```
import dt, dttest, dtStat, dtTree
# set up the learners
learners = []
me_set = [0, 1, 5, 10, 100]
for me in me_set:
    learners.append(dtngTree.TreeLearner(minExamples=me))
# load data, split it to train and test data set
data = dt.ExampleTable("voting")
selection = dt.MakeRandomIndices2(data, 0.7)
train_data = data.select(selection, 0)
test_data = data.select(selection, 1)
# obtain and report on results
results = dtTest.learnAndTestOnTestData(learners,
    train_data, test_data, storeClassifiers = 1)
CA = dtStat.CA(results)
IS = dtStat.IS(results)
print " Ex Size CA IS"
for i in range(len(learners)):
    print "%3d %4d %5.3f %5.3f" %
        (me_set[i],
        results.classifiers[i].treesize(),
        CA[i], IS[i])
```

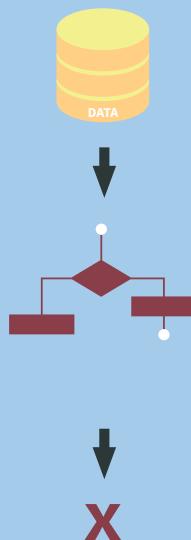
### 3.1 Machine Learning: a different type of algorithms

In order to understand the fundamental characteristic differences of ML, which we believe correspond to a paradigm shift, we compare in this section the functioning of conventional algorithms with the new type of algorithms, ML. The two major interrelated differences are, first, a new type of relationship to data and, second, the nature of the algorithm used in the production phase (vs. development one). Through this prism, we propose to compare classical algorithms, which we call here Deterministic Algorithms (DA), with Machine Learning ones.

## Development



## Production



**Figure 1:** Deterministic Algorithm (DA) in development and production phase

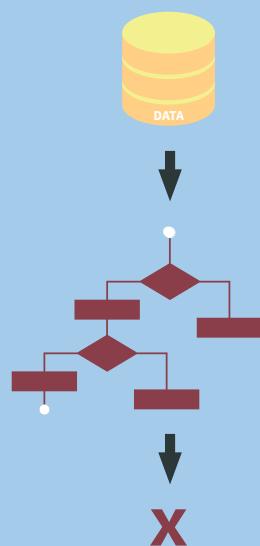
## Deterministic algorithm (DA)

Conventional algorithms usually are deterministic algorithms. Like a recipe, they consist of a hard-coded set of rules which always produce the same output. The software engineer explicitly programs the algorithm's logic without using any data. When the algorithm is put into production, data are fed to the algorithm in order to produce results. Data has no impact on the algorithm in itself.

## Development



## Production



**Figure 2:** Machine Learning (ML) algorithm in development and production phase

## Machine Learning (ML)

In contrast to deterministic algorithms, when “programming” Machine Learning we have two different phases. The first one is programming the ML algorithm itself, which is de facto what we just described for the Deterministic Algorithms. In a second phase, usually called “training”, a data scientist (or data engineer) uses the ML algorithm together with data to produce a new algorithm: the production algorithm. Often, the ML algorithm and the production algorithm get confused. Data scientist call the latter a “trained algorithm” which contains thousands of parameters that were not explicitly programmed by a human, but rather automatically “learned”, i.e. estimated, using data samples. Here, data is grown into an algorithm.

Some recent Machine Learning algorithms are given the capacity to “re-train” themselves, and as such show a certain degree of autonomy. This capacity brings additional risks, such as the danger of external manipulation. In this white paper, we do not address this type of algorithms, since it would add unnecessary complexity to our purpose.

```
import random
import numpy as np
import matplotlib.pyplot as plt
from sklearn import datasets
from mla.kmeans import KMeans
from mla.gaussian_mixture import GaussianMixture

random.seed(1)
np.random.seed(6)

def make_clusters(skew=True, *arg, **kwargs):
    X, y = datasets.make_blobs(*arg, **kwargs)
    if skew:
        nrow = X.shape[1]
        for i in np.unique(y):
            X[y == i] = X[y == i].dot(np.random.rand((nrow, nrow)) - 0.5)
    return X, y

def KMeans_and_GMM(K):
    COLOR = 'bgrcmyk'
```



## 3.2 Machine Learning and personal data: introducing GDPR

In the EU, the use and protection of personal data is regulated by the General Data Protection Regulation (GDPR)<sup>1</sup>. It is directly applicable for all companies processing data of European citizens (GDPR, Article 3) [1].

Prior to the GDPR, the regulation of personal data was based on Directive 95/46/CE. Since several notions around data privacy were considered as complex with a potential far reaching impact, the Directive introduced (in Article 29) a committee made up of representatives of the European supervisory authorities. One purpose of this committee was to provide interpretations and guidelines on the application of the Directive (Article 29 Working Party “WP29”). The WP29 is now replaced by the European Protection Data Board.

The opinions provided by the WP29 are not legally binding, but since they express the opinion of the European supervisory authorities, it is strongly recommended to follow these recommendations and to implement the appropriate measures. They feed the discussion here since they clarify what is understood or expected by several of the definitions, such as transparency.

As previously described, there is a strong relationship between data and algorithm. As we will see, the identified challenges are partially addressed by the GDPR since personal data is being processed. However, we will also see that this is only one of the issues raised by ML algorithms.

---

<sup>1</sup> Personal data are defined as “any information relating to an identified or identifiable natural person; an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identify of that natural person”



## 4. What are the intrinsic challenges of Machine Learning?

In this chapter, we want to outline the new challenges introduced by the specific characteristics of ML algorithms. When addressing the topic of regulating ML, there is a danger to drift into the topic of ethical usage of ML. Even though this important topic is closely related to our subject, we believe it should be distinguished insofar as it is not linked to the intrinsic nature of the algorithm and is already well covered by insurance law or criminal law.

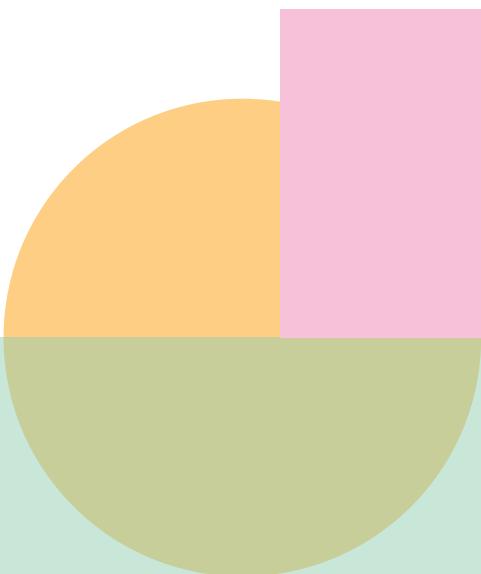
In fact, we do not cover the use of algorithms for the wrong purposes or for non-ethical intentions in this document. Even if we acknowledge that the availability and the scaling effect of algorithms are strong drivers for misuse, we consider that this ethical issue is not intrinsic to the algorithms themselves, and thus falls outside the scope of this study.

In the similar way, as we previously saw, most of the algorithms need some input (i.e. data) to execute. The nature of this input, e.g. usage of gender, age, religion, may provoke ethical questions. These ethical challenges are not intrinsic to ML, because even if they were used by a human the ethical questions remain. Thus, we do not further develop this aspect in this document.

 **Illustration:** Should a particular Machine Learning-based credit assessment algorithm make usage of variables such as specific DNA markers?

---

In the following, we focus on three intrinsic challenges introduced by algorithms based on ML: fairness, reliability, explainability. This short list strongly echoes analyses presented in several reports recently published by institutions and governmental agencies [2] [3] [4] [5] [6] [7]. For each challenge, we describe the intended objective, the problems, and the legal context. Finally, every chapter concludes with a discussion.



## **4.1 Fairness<sup>2</sup>: learning and reproducing bias**

### **What is the objective?**

We want to avoid systemic discrimination of unprivileged, protected groups. Here, we focus on the challenge of detecting and mitigating potential discrimination due to unwanted bias in ML algorithms.

### **What are the problems?**

As stated before, ML algorithms are strongly dependent on the data they use to create the “production algorithm”. Because algorithms have the potential to get deployed at scale, even a minimal systematic error will lead to discrimination of a group. In data science, this kind of error is called “bias” and may result from the process of how the data was collected. If data contains such a bias, a ML algorithm using these data for training will learn and enforce the bias.

---

<sup>2</sup> Fairness in GDPR means something different. In fact, it is stated that “personal data shall be processed lawfully, fairly and in a transparent manner” (Art. 5 GDPR). In this context, fairness is very linked to transparency and obligation imposed to the data controller related to information to be given to data subject and the respect of the purpose announced (§39 GDPR).

**Illustration:** A ML algorithm could be used to give health recommendations. In order to build the system, a detailed written survey is sent to a group of people. The collected data is used to train the production algorithm. If the survey was only available in one specific language, people not speaking this language will be prevented from participating. Without their answers, the survey will not represent their views, and the ML algorithm trained on the data will not take into account their characteristics. This group is then not properly served by the algorithm and thus could be potentially discriminated.

---

In addition to a non-representative data set which does not reflect the real distribution, the bias may have other causes. It may occur when the algorithm is used in an environment for which it was not trained in the first place. For example, if it is applied in a different geographical region or on a different group of people.

Furthermore, if the training data contain human judgment, it may have been labelled with prejudice. Since the labels serve as ground truth, the algorithm's accuracy directly depends on them. If the labels are not objective observations, as for instance coming from a measuring device, but involve human assessment instead, they can contain bias.

**Illustration:** A ML algorithm for claim approvals could be, in theory, trained from previous cases manually decided by one specific case handler. Those former decisions serve as foundation for future decisions. If this person took prejudiced decisions disfavouring a specific group of people, the “production algorithm” will replicate this bias.

Finally, ML algorithms identify pattern in data. Its major strength is the desired capability to find and discriminate classes in training data, and to use those insights to make predictions for new, unseen data sets. In the era of Big Data, a lot of data is available with all sorts of variables. When using a large amount of data, it clearly contains many correlations. However, not all correlations imply causality, because no matter how large the dataset is, it still only remains a snapshot of reality. As we will see in the following paragraphs this is a major risk in Machine Learning usage.

### **What do we know? Legal context**

At European level, several texts regulate the use of information on people in order to fight discrimination. This principle is stated in the Convention for the Protection of Human Rights and Fundamental Freedoms [8] in article 14 entitled “Prohibition of discrimination”. It is also contained in the Charter of Fundamental Rights of the European Union [9] which states in Article 21 that “[a]ny discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.”

These articles materialize the fact that society, via this regulation, expects that whatever is necessary will be done to avoid any type of discrimination. Regulation goes one level further and gives some advice on how this could be achieved by proposing to forbid the use or the consideration of some type of data. For instance, in the field of insurance and financial services, it is forbidden to use sex as a factor in the calculation of premiums and benefits if it results in differences in individuals’ premiums and benefits [10].

As a general principle, the GDPR prohibits the use of data which are considered sensitive and subject to special protection. For instance, data concerning health, a natural person’s sex

life, or sexual orientation (Art. 9) and data related to criminal convictions and offences (Art. 10) can only be processed under certain conditions (for example requiring consent of the data subject).

A processing for profiling may reveal some inferred sensitive data from correlations. In this case, the WP29 [11] recommends checking that:

- ✓ the processing is not incompatible with the original purpose;
- ✓ they have identified a lawful basis for the processing of the special category data; and
- ✓ they inform the data subject about the processing.

It is also important to note that in certain sectors, such as insurance law, specific rules exist that allow people's characteristics, such as age or health status, to be considered in order to offer them different products or services and therefore to process protected data.

For instance, in France, the regulation of life insurance allows the insurer to ask the subscriber to complete a medical questionnaire [12] that will determine if the insurer assures without special conditions, with exclusions, with a surcharge or even refuses to insure.

The French supervisory authority (Commission Nationale de l'Informatique et des Libertés) has issued simplified standards dedicated to the insurance sector [13], which determine for each purpose which data can be collected and processed. These standards specifically allow the collection of health data for contract subscription and contract management as these data will be needed to assess risk or harm. Even though these simplified standards are no longer in effect with the entry into force of GDPR, they may still be useful as guidelines.

## Discussion

The idea of preventing algorithms from unfair use of protected attributes by forbidding to use them in the training process is also known as “fairness through unawareness” [14]. However, it falls short in the case of Big Data where other attributes or a complex combination of them may serve as proxy of a protected attribute. Seemingly insignificant attributes, or several attributes combined, may provide an unexpected link to sensitive information.



**Illustration:** The information of the car model or its colour may be correlated with the owner's gender. The zip code may be correlated with the customer's race.

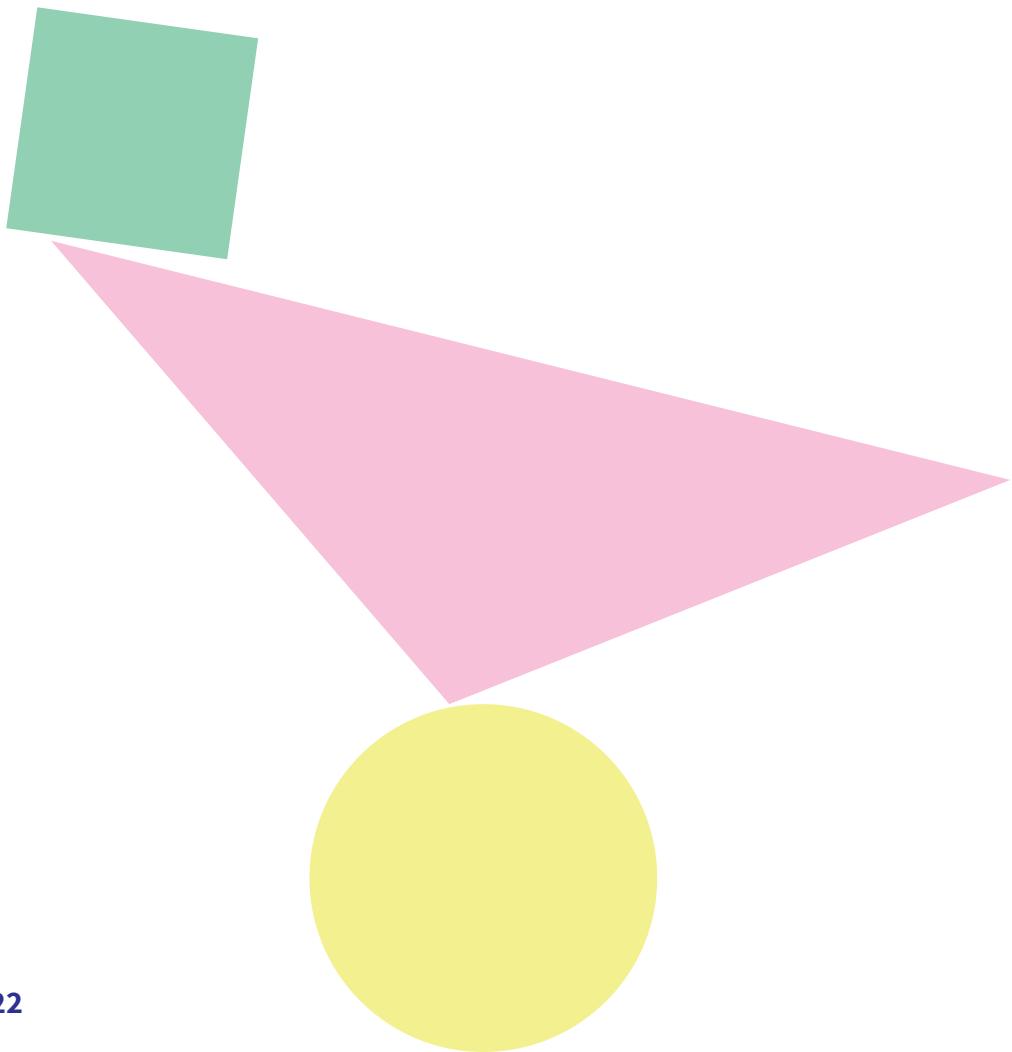
---

This risk is not totally solved but mitigated by the principle of data minimization according to which the data controller must collect and process only the personal data necessary for the intended purpose. By limiting the number of variables used, we theoretically limit the risks of finding proxies of a protected attribute. But market evolution and usage of new data, seeking for a more direct grasp of the risk, such as the one coming from connected objects (e.g. cars, home), will reveal the above-mentioned challenge.

Moreover, paradoxically, by forbidding to collect protected attributes, there is no possibility to measure for potential discrimination at a later point, which may even impede the pursuit of fairness.

From a technical point of view, in order to avoid unwanted bias as described above, we should try to detect and mitigate it by making the use of open source libraries such as “AI Fairness 360” [15], an integral part of our development workflow. Using such toolkits, we can identify bias in the training data and pre-process them if required. We can also counteract biased learning while training the ML algorithm. However, not all testing can be

automated. Defining and characterizing the protected groups needs to be done case-by-case. Also, in order to be able to measure and improve fairness, we would have to agree on a statistical definition of fairness as baseline, which is not the case today. In current research, there exist plenty of different definitions which are mutually incompatible. [16]



## 4.2 Reliability

### What is the objective?

The ML algorithms used in production need to be robust. They have to be able to cope with erroneous or unexpected input and must produce stable, predictable results.

### What are the problems?

When ML algorithms for high-stakes decision-making get deployed at scale in production, it may fundamentally impact people's lives. Therefore, reliable and predictable functioning is required for all sorts of parties. Regulators demand proof of compliance with the rules. Customers expect consistent treatment. Insurance companies need processes to ensure quality and avoid adverse selection which would put them at a disadvantage. As explained in the following discussion, achieving this objective is not always possible, even with highly accurate data.

### What do we know? Legal context

Regulation at European level, mainly the GDPR, frame the decision-making conditions, in particular on profiling. Profiling is defined as “any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyze or predict aspects concerning the data subject’s performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her” (GDPR §71).

In the case of profiling, the text insists on the need for the controller to ensure the reliability of the processing. The controller must take appropriate safeguards and 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 risks of errors is minimised” (GDPR §71 and Art. 22).

As just mentioned, one of the potential reasons of lack of reliability is the quality of the data used. The question of data quality is dealt in the GDPR in the provisions stipulating that the data collected and processed must be maintained accurate and up to date. Article 5, which establishes the principles relating to processing of personal data, states that “personal data shall be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay”.

The data controller shall regularly check the accuracy of the personal data, propose to the data subject to control their data and to correct them, and trace the changes made.

Moreover, an extra safeguard has been included in the GDPR: The regulation also states that the data subject has “the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her” (GDPR Art. 22). This means that a decision with a significant impact on a person should not be the result of an algorithmic analysis only. This decision must, as a minimum, be controlled by a human. The data subject must have the right to obtain human intervention on the part of the controller, to express his or her point of view, and to contest the decision (GDPR Art.22, §3).

## **Discussion**

ML algorithms are prone to unreliable behaviour, and its absence cannot be completely guaranteed from a technical point of view. As for conventional algorithms, constant testing should

help to detect and minimize errors. However, unwanted bias as described in the Fairness section is a possible hidden source of unreliability. It may originate from already biased, incomplete or poor training data. For rare or sensitive observations, it may also be hard to obtain a sufficient amount of training data.



**Illustration:** A ML algorithm could be deployed to predict the effectiveness of a medical product on humans. Therefore, the algorithm would be trained on data sets which were collected from test persons. However, a medical product may work differently on people of different ethnical origins. If clinical tests of the product lack big enough test groups of all ethnical groups, the medicine may have unexpected effects and hence work less reliably for people which belong to an underrepresented minority group. [17]

---

The focus on **data accuracy** in the law has limited impact. It is important when a person's data is being processed in production phase. When the input data is wrong for any reason, the result of the algorithm will be inaccurate. However, ML algorithms are using millions of data points for training, the personal data explicitly provided by one single client is just a minor part of it. The greater challenge than outdated information is the unintended and possibly undetected bias introduced by the training data (see Fairness section). A biased algorithm will certainly produce unreliable results.

Additionally, recent research has shown that complex ML algorithms may be easily tricked. So-called “Generative Adversarial Networks” create synthesised data samples which fool the ML algorithm into producing unexpected and unwanted output. [18] The differences of the data sample compared to real data samples are minimal and for humans often hard to detect. Those “adversarial attacks” demonstrate impressively the vulnerability of ML algorithms to malicious actors.

**Illustration:** In the future, an ML algorithm for claim handling is applied to assess the repair cost of a damaged car based on photos. A small perturbation – not visible to human eye – in the images could potentially lead to a repair refusal.

The requested **human intervention** and the possibility to object the result heavily depend on the transparency and explainability of the algorithm (see next section). In fact, the customer may argue, and the human controller can act or override the algorithms decision, only if they understand what the algorithm did and why.

Even more, not everything done by algorithms can be done by humans. For instance, in order to price the risk, at minima, it is necessary to consider a large amount of claim history, compare it with the customer who asked for human, and properly mutualize the risk. In our understanding, this is not obvious or even feasible to be done by a human – though this is initially defined by actuaries on large datasets with the help of algorithms.

## 4.3 Explainability

### What is the objective?

There are multiple definitions of “transparency in ML”. A weaker one is about sharing the technical details such as the source code, the variables and the training data of a ML algorithm. A stronger one, also utilised by GDPR, includes explainability, which means providing meaningful explanation of the decision-making. We are convinced that the first definition is not enough and therefore thrive for the stronger one.

### What are the problems?

Trustworthy new technologies are based on transparent processes and offer the possibility to explain at any point why a specific decision was taken. It is not necessarily expected to give insight on every single step at any time, but a human-readable explanation should be available.

An obvious obstacle towards more transparency are trade secrets.

 **Illustration:** If an insurance company knows its competitor's pricing model it can either use it to create their own products or use it to counter the offer and “steal” specific customers by playing with their pricing model.

---

The technical aspect is more complicated to solve than the business related one. Already for conventional algorithms it may be hard to give a sharp, intuitive explanation of their behaviour due to a vast number of possible combinations. For ML algorithms, this is even more difficult to achieve: ML algorithms

may have thousands of variables which are adjusted during the automated training process. The result is a new, complex production algorithm which works without human intervention. Humans provide the input data and observe the output, however, even for experts in the field it is impossible to understand the reasons why the “black box” takes a decision. The problem is not just the large number of variables, but the fact that it is usually not possible to interpret and explain the role a specific variable plays in the final decision. The outcome depends on complex interrelations of all variables.



**Illustration:** A ML algorithm is used for scoring risk of default. The insurance company will insure the loan if the score is 0.8 or higher. Customers with scores of slightly below 0.8 might ask for explanation of how the score was composed. More specifically, they might want to learn how to improve their score in order to be insured. If advanced ML is used – as of today – there is no obvious answer.

---

Moreover, it is crucial to provide some interpretability of automated decision-making: Data scientists want to be able to compare and benchmark algorithms. Regulators require to audit the algorithms in order to ensure compliance with the rules. Companies must remain in control of their processes, protect them from human attack, and provide their clients with information about their decision-making. Finally, respecting social norms and preventing discriminatory outcomes is essential because failure to do so might result in a general loss of trust in AI in the society.

### **What do we know? Legal context**

The GDPR states that “[i]t should be transparent to natural persons that personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal

data are or will be processed. The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used.” (§39 and Art. 12 GDPR).

Beyond the information on the processing and its characteristics, the data controller will have to provide the data subject with information on how the decision was made.

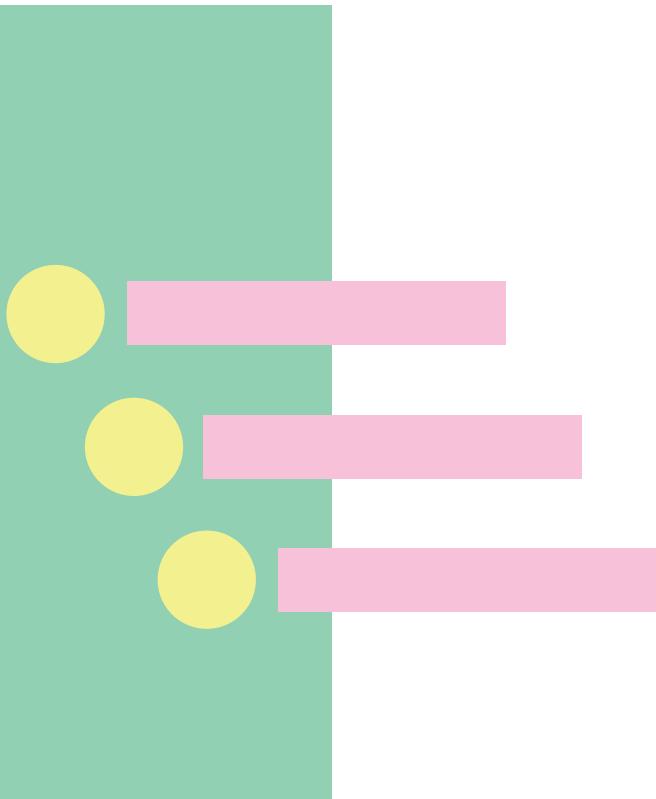
Regarding the specific case of profiling, WP29 [11] recommends that “[i]nstead of providing a complex mathematical explanation about how algorithms or machine-learning work, the controller should consider using clear and comprehensive ways to deliver the information to the data subject, for example:

- the categories of data that have been or will be used in the profiling or decision-making process;
- why these categories are considered pertinent;
- how any profile used in the automated decision-making process is built, including any statistics used in the analysis;
- why this profile is relevant to the automated decision-making process; and
- how it is used for a decision concerning the data subject.”

## Discussion

The mere notification about the fact that the customer’s data is collected and processed by an algorithm is indeed important, but not sufficient. Algorithms are part of most of the processes today. Moreover, precisely informing “to what extent” data is used in a ML algorithm is difficult from a technical point of view: E.g. personal data may be inferred from other data; or, as stated above, a decision is derived from very many variables, and the weight of a specific personal information may depend on a complex combination of those. Also, the means to explain the processes in a “black box” algorithm are still very limited. [19] [20] [21] [22]

The WP29 guidelines recommend explicitly naming the variables (“categories”) in a ML algorithm. This may not help increase transparency because a ML algorithm uses thousands of variables which get automatically learned from training data. They do not stand for separable factors in the decision-making process, and their pertinence cannot be easily explained. Communicating the classes (“profiles”) may help the customer to understand how he or she is perceived by the data controller. However, explaining why the customer was assigned this class remains hard. Also, in the case of scoring, or when the outcome of a decision process is just yes or no, this recommendation does not improve transparency.



## 5. Concluding remarks

Artificial Intelligence, and in particular Machine Learning, provide great opportunities – not only economical ones, but also in the ethical sphere itself. Algorithms have the potential to be more impartial than conventional processes based on human judgement. However, if these algorithms have hidden defaults, there is a risk of unwanted effects at scale.

All new technologies bear both beneficial potential and unexplored risks. Thoroughly understanding advantages and drawbacks before making use of such new opportunities in production is key to strike the right balance for the best of interests.

Current regulation, and in particular GDPR, is an excellent safeguard which brings the first answers to the problems. However, in the lights of the insights exposed in this document, and since the regulation addresses only the part of Machine Learning issues related to (personal) data, we believe that it is necessary to think about and work on the technical implementation of what is behind the regulations – which may include possible evolutions of ML.

Moreover, we expect that this document made clear that the described challenges are complex, and no immediate solutions exist. We would like to insist on the importance of weighting benefits vs. risks when considering any implementation.

**Illustration:** If an ML algorithm is able to predict early stages of a treatable cancer with an extremely high accuracy – but it is a black box, and in some rare cases it makes major mistakes. Maybe, the benefit of using it is higher than the associated disadvantages?

As new technologies emerge, benefits are often clear, but risks may only be fully understood at a later point. Therefore, we recommend that leaders, designers and developers of Machine Learning algorithms are not only aware of how the challenges are covered by the current regulation but also the potential negative impact of their systems.

Until responsible AI is a reality, the AXA REV R&D teams will keep working, first, on – albeit incomplete – mitigation tools for known challenges, and then on innovative algorithmic solutions for the ones without response today. If you wish to know more, do not hesitate in contacting us. We expect the broader AXA community will also help to move the needle on this highly challenging topic.

# Experts



**Boris RUF** [boris.ruf@axa.com](mailto:boris.ruf@axa.com)

Research Data Scientist



**Marie HIROT** [marie.hirot@axa.com](mailto:marie.hirot@axa.com)

Data Protection Specialist

# Sponsors



**Marcin DETYNIECKI** [marcin.detyniecki@axa.com](mailto:marcin.detyniecki@axa.com)

Chief Data Scientist



**Nicolas SHIRE** [nicolas.shire@axa.com](mailto:nicolas.shire@axa.com)

Head of Data Management



**Roland SCHARRER** [roland.scharrer@axa.com](mailto:roland.scharrer@axa.com)

Chief Technology Innovation Officer

## 6. References

- [1] “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC”.
- [2] IEEE, “The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems,” December 2017.
- [3] CNIL, “Comment permettre à l’Homme de garder la main ?, Les enjeux éthiques des algorithmes et de l’intelligence artificielle,” December 2017.
- [4] UK Cabinet Office, “Data Science Ethical Framework,” May 2016.
- [5] C. Villani, M. Schoenauer, Y. Bonnet, C. Berthet, A.-C. Cornut, F. Levin and B. Rondepierre, “Donner un sens à l’intelligence artificielle,” March 2018.
- [6] C. B. Frey and M. Osborne, “The Future of Employment: How Susceptible Are Jobs to Computerisation?,” vol. 114, 2013.
- [7] F. Maia Alexandre, “The Legal Status of Artificially Intelligent Robots: Personhood, Taxation and Control.,” *SSRN Electronic Journal*, 2017.
- [8] Council of Europe, Convention for the Protection of Human Rights and Fundamental Freedoms.
- [9] Council of Europe, Charter of fundamental rights of the European Union (2012/C 326/02).

- [10] Council of Europe, Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, Article 5.
- [11] WP29, Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679.
- [12] French Insurance Code, Article L.113-2.
- [13] CNIL, Delib. n°2013-212 concerning automated processing of personal data relating to the execution, management and enforcement of contracts implemented by insurance, capitalization, reinsurance, insurance assistance and through their intermediaries.
- [14] D. Pedreshi, S. Ruggieri and F. Turini, “Discrimination-aware Data Mining,” in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, Nevada, USA, 2008.
- [15] R. K. E. Bellamy, K. Dey and H. Michael, “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” 2018.
- [16] S. Venkatasubramanian, A. Friedler and C. Scheidegger, “On the (im)possibility of fairness,” *CoRR*, vol. abs/1609.07236, 2016.
- [17] S. S. Oh, J. Galanter and N. Thakur, “Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled.,” *PLoS Med* 12, vol. 12, 2015.
- [18] I. Goodfellow and J. Pouget-Abadie, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems* 27, 2014.
- [19] T. Laugel, X. Renard, M.-J. Lesot, C. Marsala and M. Detyniecki, “Defining Locality for Surrogates in Post-hoc Interpretability,” in *ICML Workshop on Human Interpretability in Machine Learning (Whi)*, 2018.

- [20] Z. C. Lipton, “The Mythos of Model Interpretability,” in *ICML Workshop on Human Interpretability in Machine Learning (Whi)*, 2016.
- [21] F. Doshi-Velez and B. Kim, “ Towards A Rigorous Science of Interpretable Machine Learning,” 2018.
- [22] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi and F. Giannotti, “A Survey Of Methods For Explaining Black Box Models,” *CoRR*, 2018.
- [23] D. Lee, “Tay: Microsoft issues apology over racist chatbot fiasco,” BBC, [Online]. Available: <https://www.bbc.com/news/technology-35902104>. [Accessed 24 July 2018].
- [24] C. B. Wrenn, “The Internet Encyclopedia of Philosophy,” [Online]. Available: <https://www.iep.utm.edu/>. [Accessed 13 September 2018].
- [25] Merriam-Webster, “Ethic,” [Online]. Available: [www.merriam-webster.com/dictionary/ethic](http://www.merriam-webster.com/dictionary/ethic). [Accessed 13 September 2018].
- [26] J. Deigh, *An Introduction to Ethics*, Cambridge: Cambridge University Press, 2010.
- [27] K. Shaver, “Female dummy makes her mark on male-dominated crash tests,” [Online]. Available: [https://www.washingtonpost.com/local/trafficandcommuting/female-dummy-makes-her-mark-on-male-dominated-crash-tests/2012/03/07/gIQANBLjaS\\_story.html](https://www.washingtonpost.com/local/trafficandcommuting/female-dummy-makes-her-mark-on-male-dominated-crash-tests/2012/03/07/gIQANBLjaS_story.html).

