WHITE PAPER
Banking & Auditablity Working Group
Hub France IA

Risk Control of Artificial Intelligence Systems

October 2022

BNP PARIBAS

SOCIETE GENERALE

HUB FRANCE .IA

RISK CONTROL OF ARTIFICIAL INTELLIGENCE SYSTEMS

WORKING GROUP - BANKING AND AUDITABILITY - HUB FRANCE IA

CONTENTS

1 INTRODUCTION

The success of Artificial Intelligence (AI) and the expansion of its uses in most industrial and scientific fields are naturally accompanied by the **emergence of new risks**. In this context, the definition or strengthening of regulations governing AI is developing in many regions, particularly at the level of the European Union, to promote the **controlled development** of these techniques.

The Hub France IA **Banking and Auditability Working Group**, which brings together AI and audit experts from three major French banks, BNP Paribas, La Banque Postale and Société Générale, wishes to share its thoughts and feedback on AI risk management.

This work proposes solutions to certain key issues and is positioned as a best practice **guide** for assessing and controlling the risks of AI-based solutions. The AI process will be covered from start to finish, thanks to the cross-views of the three lines of defence represented in this working group.

This organisation into **lines of defence** is specific to the financial sector, which is also particularly regulated, especially in terms of model risk management. This line-of-defence organisation makes it possible to structure the institutions' approach to risk management, through a proven organisational structure and framework. The application of this approach in the case of AI can therefore provide food for thought for other economic sectors. Beyond the organisation, the main contribution of this work probably lies in the operational implementation of the control systems envisaged. They are designed here based on expert opinion, and not as an explicit response to any of the regulations in the pipeline.

The approach adopted here consists first in describing the AI process as a whole, including the aspects relating to its compliance. In a second section, the specific risks brought or exacerbated by AI have been identified and illustrated. Finally, in a third step, ten risks were selected, based both on their importance and their specific link with AI. For each of them, proposals for impact assessment and remediation have been described.

The work does not aim to be exhaustive, which would have required a much longer document, but aims to provide a **methodological framework and best practices**. This work, we hope, will be useful to other economic sectors which, drawing inspiration from it, will be able to implement processes adapted to their context to better control the risks of the AI solutions that they deploy or use.

1 THE AI PROCESS

1.1 DEFINITION OF AI

Artificial Intelligence is a "set of theories and techniques implemented to create machines capable of simulating human intelligence". It includes two main families: **symbolic AI and digital AI**, which have each experienced periods of success and "winters", as shown in the diagram below representing the activity of these two families since the 1950s.

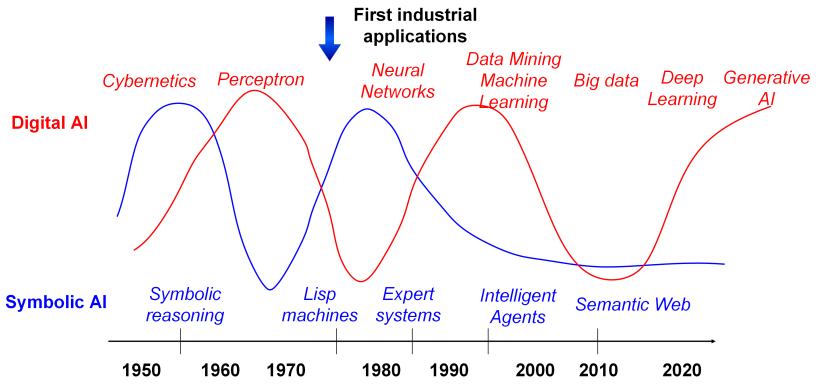

Figure 1 – Symbolic AI and Digital AI

We will not discuss symbolic AI, which is rarely used in the banking industry. We will focus on **digital AI**. Since 2012 and the success of Deep Learning techniques at the *ImageNet*[1] conference, the most widespread techniques today, and particularly in banking applications, are **Machine Learning** techniques (or automatic learning), which includes Deep Learning. In this white paper, when AI is mentioned, it will therefore refer to Machine Learning, unless explicitly stated otherwise.

1.2 AI IN BANKING

Artificial intelligence is increasingly integrated into banking processes, the trend having strongly accelerated in recent years.

A significant part of AI use cases within banks aims at **automating internal processes** to improve efficiency while reducing operational risk. Use cases are developed, for example, to automatically read and process documents in the legal field or at client on-boarding. Other uses are dedicated to the classification of incoming e-mails and the generation of a response, thus reducing the processing time of customer requests. AI also facilitates the extraction and structuring of large volumes of data, for example to reduce the workload of credit analysts.

The use of AI also aims to improve **customer experience**, with the emergence of conversational agents that assist customers in their operations (Chatbots, Voicebots), sometimes called *selfcare*. In marketing field, AI allows to refine the knowledge of customers to offer them more personalised products and pricing conditions, by a better anticipation of their needs and the

[1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems pp 1097-1105. 2012. papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

prediction of their attrition. In addition, AI facilitates the emergence of new services for customers, such as investment recommendation engines (*robo-advisors*).

By improving the accuracy of algorithms, the use of AI contributes to the **mitigation of many risks**, in particular **operational risk** (e.g. detection of fraud committed against customers, detection of anomalies in data), **credit risk** (e.g. detection of the most risky customers when granting credit) and **compliance** (e.g. in the context of the fight against money laundering and the financing of terrorism, the detection of "negative news" about customers as part of KYC[2]). There are also applications in financial models and the management of these risks, as well as in ALM (*Asset and Liability Management*).

Finally, it is important to remember that the assessment of the relevance of an AI solution must be carried out with regard to the institution's risk appetite. This assessment of relevance must also be carried out according to all the risks reduced or increased by the use of AI, in particular in addition to model risk, operational, compliance, credit or market risks.

If for artificial intelligence there is to date no precise and globally shared framework defining all the elements related to the impacts and risks of AI systems, it should however be noted that the development and deployment of artificial intelligence systems are based on highly connected and more mature **internal frameworks:**

(i) data management and in particular data quality;
(ii) Cyber Security;
(iii) Model risk management within financial institutions.

These frameworks benefit from well-established regulations, professional practices and risk management expertise that have been taken into account in this white paper.

In addition, financial institutions are very experienced in model development, particularly in designing **controls to ensure proper modelling discipline**. Consequently, some relevant risks for artificial intelligence models (*feature selection, hyperparameter tuning, overfitting*, lack of documentation), are not so predominant in the financial service industry because of the existing culture of rigorous model development. However, it is important to pay attention to them even in financial institutions when the models are developed externally or in teams traditionally far from modelling.

1.3 MACHINE LEARNING

The production of a Machine Learning-based solution is done in two steps:

- The **Build** phase: starting from a specific need, a data scientist will collect the appropriate data to constitute a training dataset, then select a learning algorithm (most often, from an open source library). At the end of the learning process, an AI model is obtained - a program that can then be used. This program can be coded in any computer language, the most common today being python. Data used for learning is necessarily data from the past. We also use a validation dataset, different from the training dataset, but having the same structure, in order to be able to compare models with each other and choose the best one. In general, we randomly cut the set of all available data into three parts, for example 70% / 15% / 15%. For learning, we use the first part to produce models, and the second to choose the best model (hyperparameters selection). The third part is reserved for testing

[2] *Know Your Customer*

the model and will never be seen during training. If too little data is available, cross-validation techniques are used.

- **Exploitation** phase (also referred to as **inference or Run**): at the end of the learning stage, a data scientist presents new data to the model obtained and gets as output the most probable result for the data entered. We therefore use data from the past to **predict behaviour** for the future. It is recommended to collect these data and the associated results over time, and compare them with what is really happening, which allows us to measure the prediction error (we say that there is an error if they are different). We can then, at the desired frequency, relaunch the training of the model by incorporating these new data. A **retraining loop** is thus set up which makes it possible to improve the model over time.

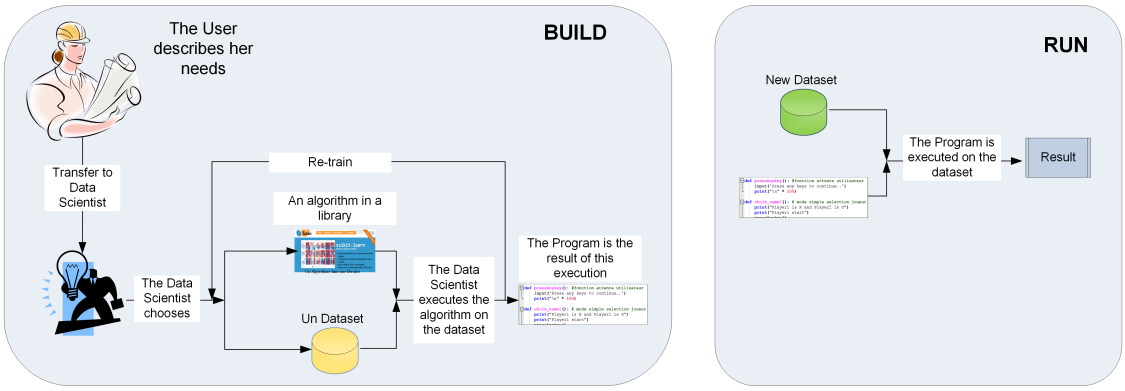

Figure 2 – The two stages of production and use of a Machine Learning model

Learning algorithms essentially correspond to two major approaches. In the first, we know the correct answer associated with a data instance and in the second, we do not know it.

These two approaches are defined more precisely as follows:

- **Supervised learning**: each data point is associated with a label, or **labelling** (a specific process called annotation may be needed to produce those labels). For classification algorithms, this is the class label (e.g., dog or cat for images). For regression algorithms, it is a numerical label (e.g., the amount of fraud on a card). Supervised learning aims at reducing the error on the training dataset. We pass the data of the dataset one after the other, then for each data, we compare the result obtained by the model with the label and we measure the error. Learning, for an algorithm, consists in iteratively testing many combinations of input data to find the output data that correspond to the true labels. The iterative approach makes it possible to gradually reduce the errors, and thus to learn. The quality of learning can strongly depend on the number of "known" events (i.e., labelled data). If we have sufficient input data, a good model can be obtained. Otherwise, another approach should be used.
- **Unsupervised learning**: data points do not have associated labels. There are many approaches, typically seeking to group similar data points together (clustering algorithms) so as to identify associations and differences between data without any a priori.

There are less prevalent approaches, such as **semi-supervised learning** or **reinforcement learning**. Vision techniques, speech recognition or natural language processing (NLP) very often use techniques based on **neural networks**, deep neural networks in general. The latter make it possible to build a **representation** of the data in a space of generally lower dimension. This is called

embedding in the representation space. Other embedding techniques are used to represent text (word2vec) or graphs (graph2vec). It is then possible, for example, to categorise documents, automatically summarise or even identify their type (ID-card, pay-slip…).

Once in production, the model is expected to behave on production data with performances comparable to those obtained during the model construction phase: the model is said to "generalise well". This property is fundamental, and it is essential to ensure that the model we use is not in a situation of **overfitting** (or over-learning). Indeed, when a model is too complex, it will be able to fully memorise (rote learning) the training dataset and will not be able to generalise correctly in production. Figure 3 below illustrates this phenomenon. When increasing the **complexity** (technically the Vapnik Chervonenkis[3] dimension) of a model (e.g., the number of neurons in a neural network, or the depth of a decision tree), the error on the training dataset decreases, along with the error on the validation set. But when the model becomes too complex, the learning error continues to decrease (the model has memorised the dataset) while the validation error starts increasing: the model no longer generalises, it "over-learns" (or "overfits") the training dataset.

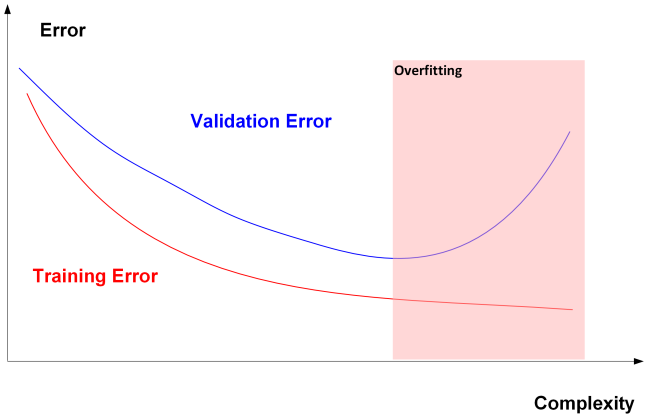

Figure 3 – Training and generalisation errors as a function of model complexity

Whatever the learning mode, the total number of algorithms is not very high (with many variants). Here is a link to a cheat sheet[4] which summarises them in five pages (dense!) or the synthesis[5].

It should be noted that **three families of performance indicators** are used in practice: **technical indicator**s are used for learning to optimise the AI model, **business indicators** are used to measure the business value generated by the use of the model, and finally **operational indicators**, such as computing time, latency, number of variables and complexity of the model, or even cost of the variables if any are purchased, are used to measure the performance of the model from an operational perspective.

1.4 MODEL PRODUCTION PROCESS

The **model production process**, shown in the figure below, includes different steps, with potentially backtracking for iteration until one is satisfied with the result. The following figure (Figure 4)

[3] Vladimir Vapnik – Estimation of Dependences based on empirical data. Springer. Information sciences and Statistics. Reprint of 1982 Edition with afterword. 2006.
[4] https://github.com/aaronwangy/Data-Science-Cheatsheet/blob/main/Data_Science_Cheatsheet.pdf
[5] https://towardsdatascience.com/overview-of-supervised-machine-learning-algorithms-a5107d036296

illustrates the actual situation with potential iterations and dependencies. The distribution of tasks between business and IT may differ between organisations.

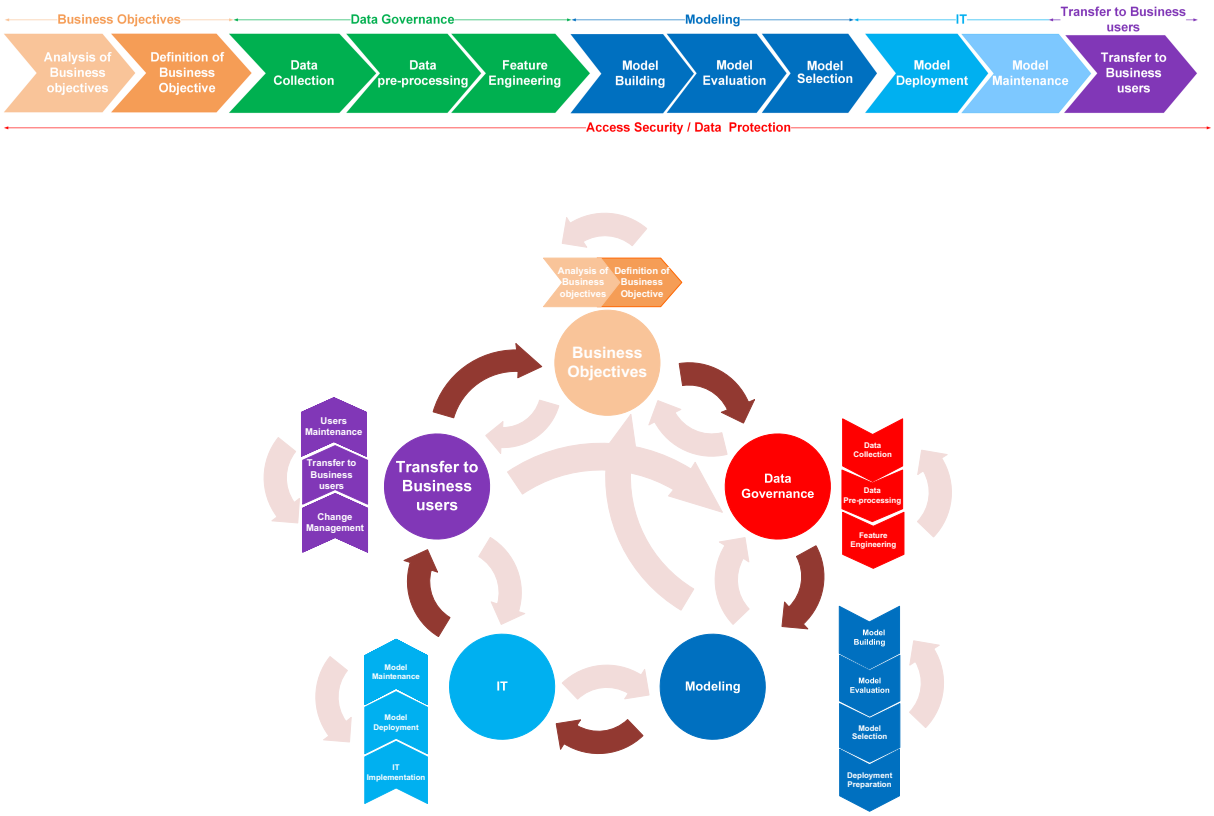

Figure 4 - Production process of a model in linear representation (top) and with iterations (bottom)

Assessing the risks of an AI solution must therefore be done **throughout the process** and it is essential to ensure that the right stakeholders are positioned wherever relevant.

1.5 BIASES

The **concept of bias** is essential to define, as it will repeatedly come up in this white paper. Biases are not specific to AI systems. However, they can be amplified or hidden by the complexity of certain models.

Bias can be defined in different ways depending on where it comes from. Historically, in the statistical sense, the bias of a model is the difference between the **expectation of an estimator $\hat{f}$ and the quantity to be estimated f**

$$Bias(\hat{f}) = \mathbb{E}(\hat{f}) - f$$

Measuring a bias is equivalent to quantifying systematic errors in a model.

Statistical bias is no longer the intended meaning when we talk about **biased data** today. This will happen if a group, for example women, is under-represented in the training dataset. A biased dataset will then produce a model that may not correctly learn the underrepresented group. The model is then likely to **produce discriminations** to the detriment of this subgroup. Biases can take different forms in data and their manipulation throughout the lifecycle of an AI system. They are therefore present at several levels:

- **Societal bias: problem of representativeness** which stems from the fact that data were collected in a specific social and historical context no longer suited to the situation in which the model is used. For example, the proportion of women among the 500 CEOs of the world's largest companies has changed a lot over the past 50 years. This limits the relevance of this information for a model intended to be applied in the current world;
- **Selection bias: problem of population representativeness** because some datasets were created from a subgroup of the global population which does not fully represent the application scope. For example, in the case of credit granting models, it is common to only use data from credits accepted and financed in the past to build the models. Thus, all rejected files are not considered in the modelling, thus creating a selection bias[6];
- **Cognitive bias** during data annotation: the annotator can reproduce a social or cultural lack of knowledge in the choice of labels to use, on images for example;
- **Association bias**: this bias can be present when variables in the model are correlated with sensitive or protected attributes which cannot be used in the model. An AI system can in some cases base its decisions on latent variables (i.e., not directly observable) representing subgroups of individuals. We also speak of **encoded bias** because the sensitive variable is encoded by the variables selected in the model;
- **Evaluation bias:** when performance is not measured on a test data set independent of the training dataset, the measurement of the generalisation power of the model is distorted (because, for example, of the use of a large proportion of the same observations between training and test data);
- **Algorithmic bias:** the choice of algorithm can also **influence predictions** because some algorithms can amplify an under/over-representativeness of a class of individuals, for example;
- **Automation bias:** relates to the fact that users could **favour results of automated systems** over those from non-automated systems, ignoring their critical thinking;
- **User interaction bias**: users will **orient future model predictions** by providing specific data, especially for continuously learning models. This is the case for many web applications[7] (**position bias, popularity bias**, etc.);
- **Feedback bias**: the use of AI results can also create a bias when a person follows this result, which is then fed back into the learning process leading to the automatic reinforcement of this result. This is a special case of **interaction bias**.

This list of biases is not exhaustive, but it does highlight that bias can impact the AI system built from the data. Furthermore, it is difficult to correct one bias independently of the others, due to their **interdependence**.

1.6 ACTORS

The development, implementation and monitoring of an AI system involves many players, some of whom are data science / AI specialists and others who are not. *Data scientists* must always have **business skills and IT skills** in addition to their expertise.

As shown in Figure 5, there are three main families of actors:

[6] Techniques such as reject inference can help mitigate selection bias by taking into account in the training dataset rejected credit applications.
[7] Baeza-Yates, R. Bias on the Web. Communications of the ACM 61, no. 6: 54-61. 2018. http://classes.eastus.cloudapp.azure.com/~barr/classes/comp495/papers/Bias-on-web.pdf

- "Business" profiles;
- Data science specialists;
- IT profiles.

Note that each bank may have different names for the actors, or even divide these roles differently among different families. The following descriptions should therefore be adapted to each case.

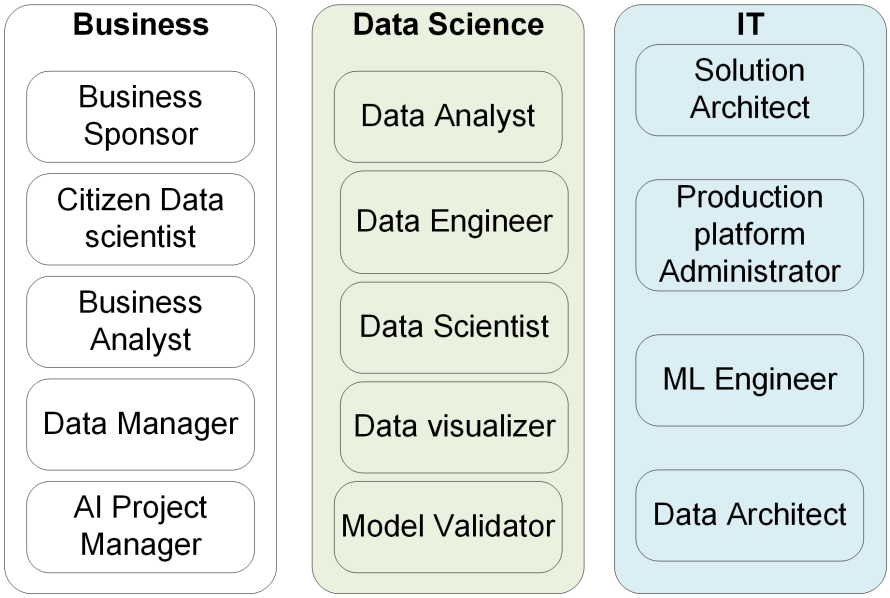

Figure 5 - Actor profiles

- **Business profiles**
  - *Business sponsor: an entity or person who expresses a specific need for a new project and who is responsible for ensuring the relevance of the project as it goes into production;*
  - *Citizen data scientist: a person that can develop data science models with simple tools (no-code) to demonstrate/explore business value; this person is not a data scientist, but rather a business-person;*
  - *Business analyst: a person who analyses business needs and identifies business KPIs (key performance indicators);*
  - *Data manager: a person who organises and manages the data in his/her area of responsibility. He/she acts as a reference for the data. Depending on the organisation, he/she can be part of the business or IT;*
  - *AI project manager: a person who leads the framing of the project, i.e., defining the functional specifications and translating into technical requirements and who also pilots the project execution. This person also ensures compliance with data regulations (GDPR, etc.).*
- **Data science profiles**
  - *Data analyst: a person that processes, exploits and analyses data;*
  - *Data engineer: a person that analyses data needs, defines collection and monitoring processes accordingly (batch, streaming);*
  - *Data scientist: an expert in data science, who designs the model development pipeline and produces candidate models and contributes to the selection discussion;*
  - *Data visualiser: a person who processes data and draws insights and develops visualisations for data scientists and business specialists;*

- *Model Validator: a person who ensures the validity of the model and evaluates its possible unwanted consequences. He/she provides an independent appraisal of the model, independent of the model developer (data scientist).*

- **IT Profiles**
  - *Solution Architect*: a person who is in charge of defining a solution architecture, including cybersecurity constraints;
  - *Production platform administrator*: a person responsible for the IT platform where the models are deployed and executed. He/she handles alerts and incident recovery;
  - *ML engineer*: a person who integrates ML pipelines (i.e., the succession of model development stages) into MLOps (Machine Learning Operations) processes;
  - *Data architect*: a person that designs the Data infrastructures and solutions and identifies the various data sources.

The **responsibilities** at the different stages of the AI system life cycle are distributed as follows:

- **Model/Product Owner**: the person who takes responsibility for the model in production (and therefore for putting it into production). This corresponds to the person on the business side who will benefit from the use of the model and who can also play the role of sponsor;
- **Model Developer (data scientist, data analyst, data visualiser**): person or group of people in charge of developing the model (modelling team);
- **System Developer**: person or group of people in charge of the IT system that will implement and then allow the model to be used;
- **Model Validator (data scientist with validation role):** person in charge of the independent review of a model;
- **Model Monitor:** person in charge of monitoring a model that is in production. This can include monitoring on the business side as well as on the modelling team side;
- **Model User:** person who uses the model in their daily work.

1.7 CONTROLS

The **purpose of a control framework** is to **manage and control risk**. Controls aim to identify risks, **qualify their impact and assess their materiality**. They may lead to the application of risk mitigation measures.

Model risk is defined and strictly controlled by banking supervisors. They have formulated **minimum requirements** for model risk management, which apply to any entity using models, depending on the type of use.

These include the following frameworks:

- (US) SR Letter 11-7, Supervisory Guidance on Model Risk Management[8], 4 April 2011, set forth by the Federal Reserve Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, which **provides guidance on model risk management**;
- (Europe) Regulation (EU) No. 575/2013 (CRR - Capital Requirements Regulation[9]) which aims to improve **transparency** on the **risks incurred** by financial institutions.

[8] https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf
[9] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32013R0575

In 2019, a high-level group of experts on artificial intelligence mandated by the European Commission, published guidelines[10] for **assessing the ethics of an artificial intelligence**. The group listed **seven criteria** ensuring **trustworthy AI**, as well as the **protection of everyone's fundamental rights.** This list of criteria and the associated questions can be used to perform a self-diagnosis of its AI.

In France, the Autorité de Contrôle Prudentiel et de Résolution (ACPR) has published a discussion paper on the **governance of artificial intelligence algorithms** in the financial sector. It proposes **four principles for evaluating AI algorithms** based on:

- Proper processing of input data;
- The performance of the algorithms;
- The stability of the model's relevance over time;
- The different degrees of explainability according to the stakeholders.

The European Commission proposed in April 2021 a **regulation establishing harmonisation rules** in the field of artificial intelligence. This regulation has a cross-cutting purpose, beyond financial institutions. It classifies[11] artificial intelligence applications according to their risks and regulates them accordingly. Low risk applications are not regulated. High-risk AIs would require mandatory self-assessment before being brought to market, and some critical applications would require independent third-party assessment. The proposal would also aim to ban certain types of AI (e.g., mass biometric surveillance or social rating). This regulation therefore introduces a **very general control framework**, depending on the level of risk of each AI. The present document does not seek to replace this control framework but proposes control mechanisms for certain AI-specific risks. This regulation is still evolving.

It should also be noted that **several definitions of AI** coexist at the level of European authorities. The Joint Research Centre of the European Commission proposes a definition based **on a taxonomic approach**[12], while the report[13] adopted by the AIDA Committee in preparation for the European Parliament's deliberation of May 2022 adopts a **more global definition** but with a less definite perimeter from a legal point of view.

[10] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai
[11] https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence
[12] Nativi, S. and De Nigris, S. AI Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework, EUR 30772 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-40325-8, JRC125952. 2021. https://data.europa.eu/doi/10.2760/376602
[13] Report on artificial intelligence in a digital age. Special Committee on Artificial Intelligence in a Digital Age. 2020/2266(INI), April 2022. https://www.europarl.europa.eu/cmsdata/246872/A9-0088_2022_EN.pdf

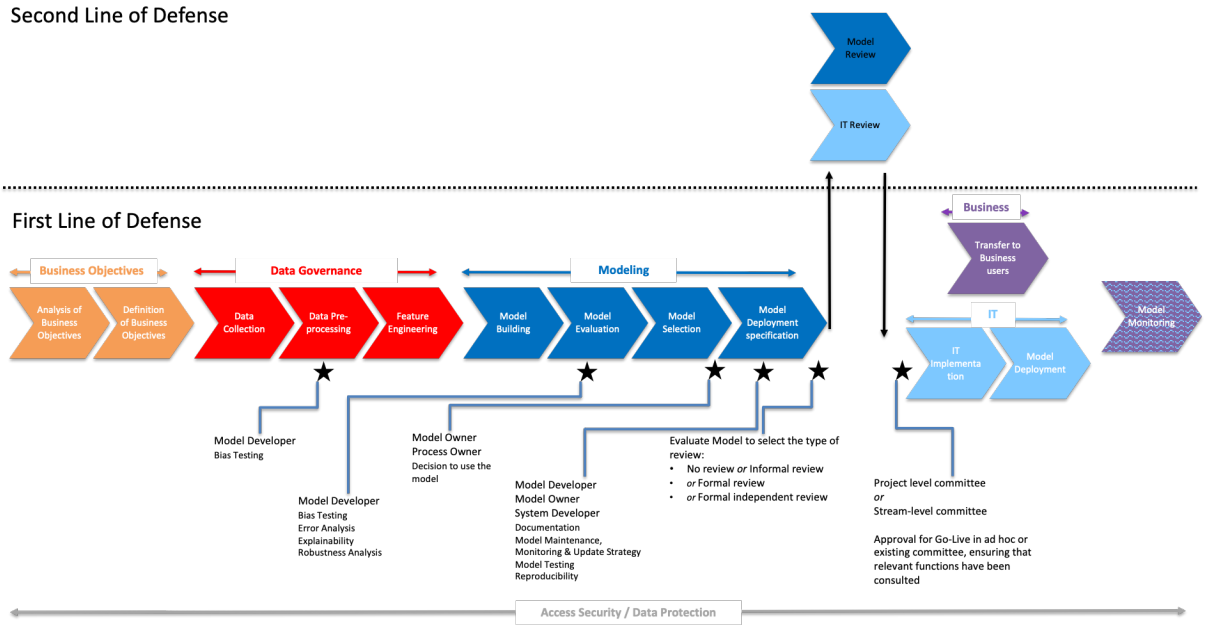

Figure 6 - AI process and lines of defence

In the diagram above (Figure 6), we have highlighted the role of the various actors in the validation process of artificial intelligence models. We present the main elements of this process at the level of the **first line of defence** (lower part of the diagram) as well as at the level of the **second line of defence** (upper part of the diagram). The second line of defence is not systematically a prerequisite for going into production.

Indeed, as it is already the case for the management of other risks in a financial institution, and with adjustments from one bank to another, the organisation of the management of model risk, and therefore of the risks that arise from the use of artificial intelligence, is organised along **three lines of defence**.

- The **first line of defence** (LoD1) is represented by the "**Model Owners**". These are the people who take responsibility for the use of models and are therefore the first to ensure that the risks associated with the model have been considered during the various stages of development, deployment, and use of the model. These individuals ensure that the development, documentation, and monitoring of the model **comply with the standards** of the associated financial institution. Depending on the type of use of the model, there may be specific procedures that stipulate more precisely the actions and controls to be taken. These specific governances are not necessarily organised by type of modelling, but rather **by use** (e.g., Credit, Insurance). In the context of AI, on the diagram above, the stars illustrate for each step the roles and aspects to be verified and documented.

  Note that the brand-new ISO Standard ISO/IEC 38507: 2022[14] of April 2022 deals with the governance implications of the use by organisations of artificial intelligence.

- The **second line of defence (LoD2)** consists in the **independent review teams**, the teams in charge of **governance and oversight of the model portfolio** and, if relevant, individuals participating in the model approval and review committees. Their role is to collectively ensure

[14] ISO/IEC 38507 : 2022. Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations; April 2022. https://www.iso.org/fr/standard/56641.html

that the first line of defence fulfils its role in managing model risk, but also to measure and "report" aggregate model risk at a defined scope. In the diagram above, the second line of defence plays an important role in **reviewing the model before the industrialisation phase**. It must cover all the risks identified and sufficiently material to justify it, whether at the level of data governance, model governance and IT implementation, and throughout the life of the model.

This is especially true in a **context of industrialisation** of use cases where risk analyses must be inserted as much as possible into the **model development** process rather than intervening downstream, potentially creating a bottleneck. In the context of AI models, it is common for this review to be accompanied by a **review of the IT solution** that will integrate the model, for example to **ensure data protection and security** or the **robustness of the system** against cyber-attacks. These reviews are not necessarily conducted by the same teams. Finally, the risk management function should keep the bank's management body informed of the assumptions used in the models, the analysis of the associated risks and any shortcomings.

- The **third line of defence** (LoD3) is the **internal audit teams** (sometimes referred to as the general inspection). Their role[15] is to assess the compliance of operations, the level of risk actually incurred, compliance with procedures and the effectiveness and appropriateness of risk identification and management systems. They may therefore have to check that the work carried out at the first two levels **complies with the rules in force within the institution**, which implies **re-evaluating the models** developed and, in some cases, challenging the controls carried out using alternative models. The third line of defence also assesses the **model risk management system**[16], the internal audit function in particular verifies the **integrity of the processes** that ensure the reliability of the institution's methods and techniques, as well as the assumptions and sources of information used for its internal models. It should also **assess the quality and use of qualitative risk identification and assessment tools** and the measures taken to mitigate risks. AI fully falls within this general framework.

2 RISK IDENTIFICATION

In order to build on our collective experience of AI in banking, each of the three banks participating in the working group started by drawing up an **inventory of the risks** incurred at each stage of the development and deployment process of an AI model. Beyond the description of the risk, we also sought to document **who was involved** (business, data scientist, user, IT or more generally the institution) and **what type of impact each risk** had (i.e., operational efficiency, financial consequences, reputational risk, regulatory risk). We also indicated for each risk the control measures that could be put in place without necessarily seeking to be exhaustive.

Once this information had been collected, we collectively reviewed the respective proposals in order to identify duplications and to refine or qualify certain risks. This **iterative methodology** enabled us to reach a **common vision of the risks,** which we describe in the following sections.

In this common review, we have tried to focus on risks that are either specific to AI systems or for which the risk is increased by the AI context. For example, information systems security or

[15] Order of 3 November 2014 on the internal control of companies in the banking, payment services and investment services sector subject to supervision by the Autorité de contrôle prudentiel et de résolution

[16] Final report on guidelines for internal governance, EBA/GL/2021/05 2 juillet 2021. https://acpr.banque-france.fr/sites/default/files/media/2021/12/07/20211207_orientations_eba-gl-2021-05.pdf

data governance are two issues that are usually dealt with globally by the company and affect AI systems.

In line with the process shown in Figure 6, we analysed the **five successive stages**: business objectives, data governance, modelling, IT and transfer to the business. For each step, we have examined the sub-tasks and associated risks. The control elements will be discussed in more detail later in the document (section 3).

2.1 BUSINESS OBJECTIVES

2.1.1 BUSINESS NEEDS ANALYSIS

AI systems are designed to **meet the specific needs of certain domains / use cases**. The first step in defining these business objectives also constitutes **the first risk factor associated with AI systems**. This is not a risk specific to AI, but it is potentially higher for several reasons that should be kept in mind so as to limit their impact.

2.1.2 UNDERSTANDING OF AI

One of the main causes for failing in defining adequate business needs is intrinsically linked to a misunderstanding of what AI is, in particular what it is really possible to do, the **reliability of results** or their **relevance**. In the design phase of an AI system, the teams impacted by this risk are the business and *data scientists*. The risk can materialise in several ways, starting with an unmet business objective, or a lower than expected financial performance.

2.1.3 CHOOSING THE TARGET VARIABLE

An AI system most often has the objective of **producing a result, called a target variable** (there may be several). If this target variable is poorly defined, in relation to the use case or context, the AI system may produce results that do not answer the need. Data scientists may be impacted because the modelling of the system will be inadequate, and business users of the results may suffer from poor performance.

In addition to the measures in the previous point, particular attention must be paid to the definition of the target variable:

- Exhaustive and shared description of the target and its measurement;

- Control of the adequacy between the target variable and the operational reality.

2.1.4 AI FOR PROCESS OPTIMISATION

In addition to the calculation of target variables or forecasting by learning, AI can also be used as an **optimisation engine** (of a portfolio, a process, or the detection of atypical cases). In this case, the output of the AI system must be **adapted to the business needs**. In particular, one must ensure that the margin of error from the AI model is compatible with the objective sought, and that the probability of not finding an acceptable solution is tolerable from a business perspective.

Here again, the impact of the risk can be financial or leading to a sub-optimal solution from the point of view of the process to be optimised, or even misleading and providing an inadequate solution to a problem. Quantifying the risk linked to the model implies doing so in relation to a purpose that is defined by the business and the projected use of the model. The definition of the use and the limits of use of the model are therefore essential to assess their compatibility with the theoretical operating hypotheses and the calibration carried out. The model risk is **measured by the potential deviation from reality**. To assess this risk, it is therefore necessary to have a tangible observable reality defined by the business on **relevant metrics**. To limit the risks, a certain number of measures can be considered, such as:

- A proper definition of the problem to be optimised and its scope;
- An analysis of the impact of theoretical forecasting or optimisation errors;
- A good understanding by the business of the risk of error and the choice of relevant "business" metrics.

2.1.5 THE MISMATCH OF BUSINESS NEEDS WITH OTHER OBLIGATIONS

The objectives of the business may be legitimate, but the use of an AI system may be contrary to the regulations, ethics or values of the company. In this case, the risk goes beyond the business - data scientist duo, as it potentially affects the company as a whole (fines, reputation, etc.).

There are many root causes, most of which are not specific to AI, such as the use of prohibited data. Most of these risks are covered by other control measures. Nevertheless, it is necessary to:

- Ensure that the implementation of an AI system has the **same level of assurance** as other projects (particularly in terms of GDPR, ethics, new product processes, security, etc.);
- Take into account the **regulations specific to the use of AI** (to date, these regulations are still being drafted: see the European Commission's Artificial Intelligence Act[16] project);
- Specifically for AI systems, test for undesirable effects, in particular undesirable behaviours that only become apparent with use in real conditions. This could be, for example, the discrimination of certain categories of customers on the basis of apparent age or residence, even if these are not explicitly part of the model calibration data;
- Putting in place a **governance** to assess whether senior management is aware of the usage.

2.2 DATA GOVERNANCE

2.2.1 DATA COLLECTION

The **data collection process** upstream of the model development is essential for the relevance and quality of AI systems. The quality of the input data to a model is an important issue, but not specific to AI-based models, as poor quality will also impact classical statistical models or reporting systems. Moreover, data quality is already subject to **specific regulations** (e.g., Article 82 of Directive 2009/138/EC of the European Parliament and of the Council in the context of Solvency II[17] for insurance, BCBS239 of the Basel Committee on Banking Supervision[18], Order of

[17] https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32009L0138&from=sv
[18] https://www.bis.org/publ/bcbs239.pdf

25 February 2021 amending the Order of 3 November 2014, Article 104[19]). The subject is therefore not specific to AI-related models.

In particular, the Basel Committee defines **14 principles divided into four themes**. The first five involve putting in place:

- **Governance** related to risk data aggregation capabilities and risk reporting practices;
- A **data architecture** and **IT infrastructure** to strengthen its risk data aggregation capabilities and risk reporting practices, not only in normal situations, but also in times of stress or crisis;
- A **requirement for accuracy and integrity** of risk data;
- A **requirement for completeness** of risk data;
- A **requirement for timeliness** in order to be able to quickly produce, aggregate and update risk data.

If the purpose of these principles is to allow for better risk measurement, it leads to an organisation within banking institutions that contributes to **a cross-functional vision of data quality**, whether it is linked to reporting, AI models or any other business purpose.

2.2.1.1 DATA AVAILABILITY

The ability to access essential data for the AI solution must be assessed. The **availability of useful data** can be found at two levels: during the **modelling phase** (build) and in **operation** (run).

During the modelling phase, availability issues arise in particular during learning, when essential data are not available, such as a truncated or incomplete history for technical or regulatory reasons. It may also happen that the **retention period** in the active database is set by the data protection officer at three years, but for which a period of at least ten years would be necessary in order to correctly model long-term processes.

During the operational phase, some data may be missing, invalid or available with a delay compared to the use of the model.

These risks can be limited by analysing the input data during the modelling phase and by continuous monitoring of the data during the operational phase.

2.2.1.2 EXTERNAL DATA

The entity responsible for a model might not necessarily control the quality of the external data. The associated risk factors stem from a **possible lack of transparency** regarding the collection and definition of the data, the measurement, aggregation, or calculation rules used. Unlike internal data, where the data producers are generally responsible for the quality, it is the responsibility of the *data scientist* or *model owner* to ensure the **quality** of the external data used.

As with data availability in general, the risk is present during the **calibration of the models** but also during the **operational phase**.

2.2.1.3 UNAUTHORISED USE OF PERSONAL DATA

Banking information systems contain a large amount of **personal or sensitive data**. As foreseen in the General Data Protection Regulation (GDPR), the collection of consent to the use and

[19] https://www.circulaires.gouv.fr/loda/id/LEGIARTI000043224879/2021-06-28/

sharing of certain data is essential in specific cases. The uses of the model must also be considered in this context, as consent is given for explicit purposes at the time of the agreement. Another risk factor lies in the use of **historical data**, taking care not to exploit data that has exceeded its regulatory archiving period or is subject to the right to be forgotten. The length of retention is a risk to be analysed specifically, with the risk of conflict between global and local regulations.

The data to be checked may concern individual variables but also **cookies and other tracers** for example.

These risks must be addressed **upstream of the constitution of data sets** and be based on strict data governance. In this respect, it is necessary to ensure that data scientists and the functions in charge of personal data governance are properly informed about the intended uses.

2.2.1.4 ANONYMISATION OF DATA

The ISO/IEC 29100: 2011[20] standard for the protection of personally identifiable information defines anonymisation as a process by which personally identifiable data is **irreversibly** transformed so that the persons concerned are no longer identifiable[21]. Pseudonymisation, on the other hand, allows identification to be re-established by means of additional information.

The **use of anonymised data**[22] for model calibration is strongly recommended, except when it impairs the relevance of the dataset, or when anonymisation is incompatible with the purpose, such as processing granular (detailed) and non-aggregated data. In this case, **pseudonymisation** techniques should at least be applied, ensuring that the consent of the individuals is appropriate for the purpose. It is therefore necessary to assess the **risk of re-identification** and the compatibility of the techniques used with the relevant regulations, for example the General Data Protection Regulation (GDPR).

2.2.1.5 DATA PROTECTION

Data protection is part of the security policies of institutions' information systems. It is naturally reinforced for personal or sensitive data. The conditions for developing and implementing AI systems can leave **security gaps** at all levels: access to models, access to calibration engines, use of external or open components, exposure of assets outside a secure environment, etc.

Modelling and deployment of an AI system must also be in line with information systems security policies.

[20] ISO - ISO/IEC 29100:2011 - Technologies de l'information — Techniques de sécurité — Cadre privé . https://www.iso.org/fr/standard/45123.html
[21] Avis 05/2014 sur les Techniques d'anonymisation. 0829/14/FR WP216, Groupe de travail « Article 29 », chapitre 2.2). https://www.cnil.fr/sites/default/files/atoms/files/wp216_fr.pdf
[22] Anonymization of personal data, CNIL, 19 mai 2020. https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

2.2.2 PRE-PROCESSING OF DATA

2.2.2.1 INCOMPLETE OR BIAISED DATA

The **nature** of the data, its **measurement** or **coding** can **bias** the information, which can distort the results when it is a target variable or, for example, contribute to a misclassification when it is an input variable.

Thus, **model input data** should be carefully assessed during the modelling and inference phases. They may be fragmented and lead to **calibration biases**, particularly due to poor sampling. They may **lack representativeness**, either in terms of historical depth or in the proportion of certain categories. Historical data may reflect previous model decisions and contain biases. Finally, they may also be the result of incorrect pre-processing (e.g., categorisation, labelling or normalisation).

2.2.3 FEATURE ENGINEERING AND FEATURE SELECTION

In the modelling process, once data have been identified, shared, protected, and analysed in terms of data quality, the next step is to build on this input data to define additional data features (***feature engineering***) but also, in a second step, to select from a wide range of candidates which data feature will be used in the model (***feature selection***). These represent therefore two stages of data manipulation which may involve related risks.

These steps are not specific to Machine Learning modelling. Maybe more specific to AI, the ability of models to ingest a large number of input variables imposes **a real modelling discipline to manage the risks.**.

2.2.3.1 USE OF SENSITIVE DATA

When choosing the data features of a model, the most obvious risk is to use data that may **induce bias or discrimination**, such as gender or age. A list of so-called sensitive data is given in the GDPR regulation. Nevertheless, *Model Owners* and *Model Developers* should be aware of the data that is not allowed as direct modelling input variables. There are some specific cases, for example in insurance, where the use of data feature such as age is accepted when it is not in general. The use of sensitive data presents a **regulatory risk** as well as a **reputational risk**.

2.2.3.2 MISINTERPRETATION OF DATA

The fragmentation of data management in organisations can also lead to a situation where the modelling team has a **poor understanding or interpretation of the data being manipulated**. Indeed, the people in charge of providing the data to the modelling team (IT) are often separated from the people with knowledge of the data (typically the *Data Managers*, *Model Owner* and *Model User*). As a result, the modelling team may end up using a variable called "Default" as the target variable, but which in fact does not correspond to the actual default variable that would be modelled, and which may be called "D_real". The consequence is that the modelling is done on an approximate basis and therefore leads to **potentially incorrect results**.

2.2.3.3 PROFUSION OF EXPLANATORY VARIABLES

Thanks to the simplified access to data and to the computational capacities that allow the learning of models with a very large number of parameters, the number of explanatory variables in a model's input is no longer a factor that limits the modelling. This makes it possible to work on **richer and more nuanced models**, but the profusion of input data is also associated with **risks** such as: (i) the increasing number of data sources which can have a significant negative effect by **raising the technical debt** and (ii) the **increasing of potential vulnerabilities** of the model in the event of failure of one of the sources.

2.2.3.4 FEEDBACK LOOPS

In supervised model learning (continuous or regular), the **use of the model in production** automatically generates additional data that allows a model to be re-trained quickly so that it can update itself as it goes along. Continuous learning generates interactions between the data and the model that can have **negative consequences**.

We illustrate **three feedback phenomena**:

- In the case where the user only gives feedback on the model proposal if the prediction is for example positive (as in fraud suspicion checks), it is obvious that the **feedback will be biased** and may lead to a **degradation of the model performance**, especially for false negatives. This is a form of **selection bias;**
- In the context of recommendation systems, for example movie recommendations on video-on-demand platforms, the model directly influences the selection of its own future data. This is not limited to recommender systems. A model for predicting real estate prices, for example, can also generate this phenomenon if economic actors use the model as the basis for their price choices. In such cases, the observed performance of the model tends to improve, but this is a **selection bias effect**. The risk is therefore to lock the model into a filter **bubble** ;
- In less frequent but possible cases, where two AI systems interact, attention must be paid to the interaction between these two models. An example is the case of **cascading models,** where the output of one model is used as input to another model. This case can occur, for example, when a first model is used to create alert scenarios to protect against data leakage and a second model is used to disqualify certain alerts to reduce false positives. It is important to be aware that modifying one of the two systems may render the second one **deficient**.

2.3 MODELLING

2.3.1 MODEL DEVELOPMENT AND SELECTION

The development phase of an AI system presents different attention points. The model's hyperparameters must be calibrated to avoid **methodological pitfalls**. It is also essential to **monitor the reinforcement of potential biases** in the model output. **Performance indicators** must be aligned with the **initial objective**. Finally, **documentation** plays a **key role** in ensuring the justification and transparency of modelling choices.

2.3.1.1 CALIBRATION OF HYPERPARAMETERS

In an AI system, hyperparameters can come from optimisation algorithms (e.g. **learning rate**), from non-convex/non-differentiable problems (e.g. choice of activation function, neural network architecture) or from statistical considerations (e.g. kernel parameters). It is important to be able to identify them and to evaluate their **potential impact on the performance** and **robustness of the AI system**.

Unlike model parameters (e.g., weights in neural networks or coefficients of a linear regression) which are learned via an optimisation procedure of an objective function related to the use case, hyperparameters are initialised using prior knowledge. They require the implementation of an **iterative process of exploration** of the space of hyperparameters. This process must be done on a **validation dataset**, different from the training data used for learning the model parameters, and different from the test data used to obtain an **unbiased measure of performance.**

The hyperparameter calibration process indirectly introduces information from the validation dataset into the model parameters, which is called **target leakage**. For hyperparameter calibration, the modelling data must be divided into two parts: **training and validation**. The independent validation data is therefore used to optimise the hyperparameters without the risk of target leakage. However, the use of the cross-validation technique also allows this problem to be addressed while retaining a larger sample of data for the final training. The k-fold cross-validation randomly divides the data into k groups of samples. In k iterations, the model is trained on the collection of k-1 groups and validated on the remaining group. The remaining group is different in each iteration. The final performance is the average of the performances observed on the remaining group during the k iterations. The final model can then be re-trained on all the data with the hyperparameters showing the best average performance.

The usual techniques for searching for optimal hyperparameters (grid search, random search, hyperband and Bayesian optimisation) are **costly in terms of computing time**. The data scientist must often balance computation time and exploration of the space of possible hyperparameter combinations. A good practice is to use **learning curves** to track performance improvements according to the chosen hyperparameters. Indeed, for numerical and critical hyperparameters such as the number of iterations (e.g., neural network, gradient boosting) or the number of estimators (e.g., random forest), it is recommended to display the two error curves according to the number of iterations for training and validation. By increasing the number of iterations, the error will continue to decrease on the training data. To avoid overfitting and to select the appropriate number of iterations, it is sufficient to select the parameters corresponding to the place where the error curve in validation starts to increase (Figure 3). This visual technique makes it easy to select the hyperparameter under consideration. The learning curves are also useful for choosing the learning rates of the optimisation algorithms and for visualising the convergence speeds of the latter.

2.3.1.2 CREATION OR AMPLIFICATION OF BIAS

Certain modelling choices, aimed at optimising performance, can **create bias or potentially amplify pre-existing bias** in the raw data.

The choice of Machine Learning algorithm can itself be a factor of bias by favouring one type of variable over another. For example, a Random Forest model will tend to give more im-

portance to continuous variables or to categorical variables with high cardinality. Some AI systems may, by design, rely on latent variables (synthesising several explanatory variables), and are thus likely to identify and exploit sub-groups of the population, without this being explicit.

When choosing a model, it is therefore important not to use performance alone, but also to **analyse the biases that may be created** and to perform a trade-off between the two criteria.

2.3.2 KPI MODEL EVALUATION AND KPI MISMATCH

Performance assessment is an essential step before deployment of the model and throughout its life cycle. It helps to avoid implementing or maintaining an underperforming model in production, thus preventing the materialisation of model risk.

To achieve this, particular attention must be paid to the **choice of the performance indicator**, which depends on the type of model (e.g., classification or score) and the intended use of the model. The chosen indicator and the associated threshold must be understood by the business in order to be able to approve the model in an informed manner, as previously mentioned in the section on business needs.

In addition, there are multiple performance indicators that can be interpreted differently for the same model.

For example, in a classification problem (positive/negative), the performance KPIs often refer to the **confusion matrix**[23]. The accuracy of the classification may be excellent, if one class is preponderant, but may not correctly reflect the performance of the model. Other metrics extracted from the confusion matrix, such as precision, recall, specificity or F1-score, should be considered depending on the nature of the problem to be solved and the business need (e.g. minimising false positives, false negatives or both).

Finally, it is important to measure the performance of the model on a sample that is independent of the one used to build the model to **identify overfitting problems**. In this case, the model performs well on the training sample, but a significant drop appears on the validation/test sample. Furthermore, the test sample must be representative of the model's scope of application in order to provide a **realistic view** of the model's performance after deployment. Then, in the deployment phase, the input data received must respect the distribution of the training data. The test sample must be representative of the application perimeter of the model. When there is a drift in the distribution, which can be detected by a drop in the business KPIs measured, the model should be retrained on a representative dataset.

2.3.3 LACK OF AUDIT TRAIL (DOCUMENTATION, LIST OF LIBRARIES USED)

Imprecise documentation of the model may question its rationale and the **reliability of the decision mechanism**. For the sake of transparency of modelling choices, documentation therefore plays a **key role**. The assumptions and the modelling choices made must be supported by **theoretical, business and experimental evidence**. It is therefore recommended that the first line of defence be trained on properly describing the elements required in the model documenta-

[23] The confusion matrix is the summary of the classification results, it compares observed and predicted data and indicates the number of well predicted observations according to the class (true (TP) and false (FP) positive / true (TN) and false (FN) negative). From this matrix, different performance metrics can be extracted, such as precision (TP/TP+FP), recall (recall or sensitivity TP / TP+FN), F1 score (2 x recall x precision / [recall + precision]).

tion. In particular, the description of the purpose and scope of use is key, along with the regulatory framework, the assumptions made, the theoretical and business support for the modelling choices, the model explanatory variables, the choice of performance metrics, the experimental conditions and results, and the model limitations. In addition, the data transformation pipeline should be documented to **ensure auditability and replicability**, particularly to **reduce operational risk** in a scenario where the people who developed the model resign. Finally, model documentation must be maintained throughout the model's life cycle.

A mismatch between the model documentation and the code is a major risk. Indeed, it is possible to question what will finally be implemented. The development and inference codes must therefore be documented and aligned with the model documentation. This will facilitate its maintenance. In order to ensure the traceability of changes, **code versioning** must be correctly applied. The data used to design the model must be archived as far as possible. Furthermore, the version of the libraries must be mentioned. Given the **inductive nature of AI**, it is key to specify the random states used for **replicability purposes**. Finally, when using third party libraries, the risk of overriding the terms and conditions of use is important. It is therefore essential to check over time that the use remains in conformity with the licence.

2.4 IT

After the model building stage, and like any IT application or service, a model must be deployed and integrated into the existing production IT ecosystem so that it can be used by the business, integrated into an existing application or as a service that can be used by several businesses or applications. It is common to see a model developed on a so-called **Data Science platform**. The model is then transferred over to IT teams in order to put it into production. This separation between the two activities can lead to a number of problems and inefficiencies.

2.4.1 DEPLOYMENT OF THE MODEL

As mentioned above, the purpose of a model is to **provide a service** (a prediction) and to be used within one or more business processes. Deployment of a model consists mainly of **preparing** it for insertion into a production IT environment. Commonly encountered activities are:

- **Preparation of the target environment** in which the model will be used (e.g., the environment of the application that will integrate the model, docker-type environment, etc.);
- **Automation of pipelines**, in particular relating to data (example: pipeline for preparing and modifying data before it is used by the model);
- **Versioning** of the model and associated metadata;
- In some cases, the model may be **recoded** for language, performance or environment issues.

After this preparation phase, deployment can be carried out, usually by the IT teams responsible for production. Several strategies (preferably defined upstream and in agreement with the business managers) can be followed, such as replacing an existing model, or using methods such as:

- **Shadow mode**: the new AI system is deployed in parallel with the current process, but its output is not used for production. This allows the stability and robustness of the performance of the new system to be assessed but requires IT resources to run both processes at the same time.

- **Canary**: the deployment is done in a gradual way on an increasing number of users or operations. In this strategy, some users test the new system in real conditions.
- **A/B Testing**: users are randomly divided into two groups A and B who will use different versions of the AI system in production. This allows empirical evaluation of which system performs better under real-world conditions based on business metrics.

This list of deployment strategies is not exhaustive. It is therefore up to the Model Owner to choose the deployment strategy that is **adapted to the risk context** of the model.

The deployment of an AI model is not only the responsibility of IT. It must be prepared as far upstream as possible because several issues may arise:

- When preparing the model for the target production environment, it is important to ensure that the **production environment will be sufficiently sized** for the model to have a satisfactory response time in relation to the number of users and "calls" to the model.
- The deployment of the model must follow the **company's security rules and production standards,** otherwise the stability and/or security of the production environment will be compromised. Some data science tools allow for rapid "production releases" by the data scientist. Depending on the company and the environment, this may be allowed according to certain criteria (e.g., demonstration or POC). In a secure production environment, the deployment of a model must follow the company's specific standards.
- In relation to a company's release and deployment standards, it is necessary to ensure through a **process and controls** that the models that are deployed correspond to the model that has been developed, tested and validated by the business, without major modifications.

2.4.1.1 USE OF THE CLOUD

The **use of the cloud** as a platform for the development and deployment of artificial intelligence models is common, particularly as cloud providers offer tools that make training and deployment tasks accessible to data scientists.

When the cloud to which data scientists have access is formally integrated into the company's IT governance, the related risks have been considered and the processes that govern the use of these resources mean that there are no particular risks specific to the use of artificial intelligence. In most cases, these clouds are either hybrid or private and offer certain guarantees in terms of data protection.

On the other hand, attention must be paid to the use of cloud resources outside this IT governance. Indeed, in university courses, teaching often relies on public clouds as sandboxes for students' practical work and projects. Young graduates therefore develop expertise in the use of these resources but are rarely made aware that in the context of a private company, certain aspects - exposure of confidential data, exposure to extraterritorial regulations, intellectual property - can be problematic. There is therefore a need to ensure that there is **some awareness** among newcomers. If it is necessary to use a public cloud for access to sufficient computing power, particularly GPUs (Graphics Processing Units), then it is important to ensure that an **encryption process** is in place. In any case, the tools used must be listed in the IT catalogue and non-referenced tools (shadow IT) must not be deployed without the agreement of IT security.

Note that there exists a regulatory environment for the cloud, including the European DORA project[24].

2.4.1.2 CYBERSECURITY OF AI

An AI system, like other IT assets, is susceptible to most cyber security attacks. The level of risk is a function of the **exposure and accessibility** of the model and the data used to train the model and is intrinsically linked to the security the company has in place to secure the IT systems.

The main types of attack specific to AI solutions are:

- Poisoning: the training data is for example modified to induce a change in the model results. This assumes that the attacker has access to the training data;
- Oracle: the attacker attempts to extract information about the AI model, or even the data used for training it, by "querying" the model a large number of times and analysing the results;
- Evasion: the attacker, leveraging knowledge of the AI model or how it was created, modifies some of the input data to the AI model in order to produce a different result than the model would have given.

In particular, **adversarial attacks** can be produced by constructing adversarial (or sometimes called adverse) examples, i.e., examples to which an imperceptible perturbation (the **adversarial perturbation**) has been added, and which, because of this perturbation, are misclassified. For example, a spam message in which a character has been modified is now classified as non-spam. The problem is to determine the **disruption** that will bring the desired result. These attacks can be implemented during **training** (poisoning) or in **production** (evasion).

2.4.2 MODEL MAINTENANCE AND MONITORING

When a model is in production and providing a service, it is necessary to ensure that this service does correspond to the expected one. This involves mixed IT and business teams. In addition to IT monitoring to ensure that the service is available. The relevance of the model results must be regularly assessed to verify the adequacy of the model usage.

- In the same way that the performance of the model was assessed during its development, it is necessary to ensure that this **performance remains the same** throughout its use. This implies defining **relevant indicators** to monitor this performance, and to be able to follow it over time. This is especially true if a model is automatically retrained to identify a possible deterioration in performance, when the ground truth is not yet available (e.g., in fraud or money laundering). One solution is to keep a sample that is systematically **analysed manually**. This implies additional costs and therefore it is important to size the sample accordingly;
- This performance monitoring must be accompanied by an operational process that deals with degradations in model performance, with the definition of **significant thresholds** *a priori*. This process should involve the team that created the model in order to determine the **causes of the performance degradation**: are the data used still representative of the desired objective? Does the algorithm need to be changed to improve performance? Is there

[24] Proposal for a Regulation of the European Parliament and of the Council on the digital operational resilience of the financial sector. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0595

a change in user behaviour? Are there any data quality issues (or a change in external data)? An analysis should be carried out to determine the cause of the decrease in performance compared to the initial model and the solutions to be applied.

In addition to the performance of the AI model, its **carbon cost** should also be assessed and documented. Thus, the **type of hardware** (e.g., GPU), the **memory used** and the **performance in terms of computing time** must be reported for the experimental phases, in particular when training the model. Indeed, the energy consumed to power the hardware and memory depends mainly on the type of hardware and the time of use. Furthermore, the efficiency of the **data centre and its location** will determine the final impact in grams of greenhouse gases emitted.

2.5 TRANSFER TO BUSINESS

The transfer of a model to the business corresponds to its **operational implementation**. This phase includes a set of elements that are not specific to AI: deployment of new tools, training, change management, support, control and monitoring systems.

However, the use of AI implies **new risks** and **particular adaptations** of the systems. The risk can be increased when the development framework of the model does not correspond to the usual project management and production framework that users are used to.

2.5.1 INTERPRETING AND EXPLAINING THE RESULTS

The way in which the AI-based model produces results is not necessarily intuitive and its operating logic may be obscure. One first **risk** lies in the **transfer** of the model experts to the business. These experts must therefore be in close contact with the business. Even if the **risk of mismatch of the solution with the actual need** has been dealt with upstream, the final result and the model's ability to meet the need may still vary. Among the **points of vigilance** to be dealt with, particularly during training, we can note:

- The use of the model by the business **process owner**: the risk lies in **an ill-advised positioning** of the AI-based brick within a business process. This risk can appear when the process owner does not have a good understanding of the purpose of the proposed model;
- The description of the type of result produced according to the **input data, the nature and the possible margin of error** on the production of the model and the ability to challenge the model prediction. Models that produce a reliability index in addition to their results moderate this risk. This risk can be mitigated by a **general acculturation to AI** as well as a more **advanced specific training of business** users, with borderline cases illustrating the margin of interpretation to be used. The risk here lies with the direct users of the model;
- The proper use of the results produced: the model has been designed for a specific purpose and context. The **risk of misuse** must be controlled, in particular the risk of using the results of the model for a purpose other than the one initially planned.

2.5.2 ROLES AND RESPONSIBILITIES

During the construction and deployment of the model, roles and responsibilities may be **poorly defined or misunderstood**, which can create a risk to the business. Among the attention points specific to AI, the Model Owner must ensure that all the following tasks have been defined and assigned:

- **Ongoing monitoring** of the model: who is responsible for monitoring the performance of the model? Which indicators are monitored? What tool is used for this monitoring? What thresholds or rules are applied to trigger alerts? Whom to alert, in which context?
- **User support**: who is the right person to contact in the event of aberrant or unexpected results: the traditional user support or the model designers?
- **User training**: who is in charge of training new users and who is responsible for maintaining know-how?
- **Post-recalibration process**: what type of validation after a model recalibration or version change (e.g., when the model exploits transversal libraries that are upgraded)?

3 CONTROL MEASURES FOR THE TOP10 RISKS

Having defined a list of risks, as described above, our working group sought to establish which risks appeared to be the most significant in terms of **associated impacts**, taking into account our experience of the issues, the controls in place and the materiality of the impact.

To do this, each of the three banks reviewed the list of risks and ranked the risks presented from 1 to 10. We then jointly reviewed our ratings, in order to challenge our assessment. The result of this **qualitative homogenisation** is presented in this section, ordered for ease of reading according to the overall steps of the process presented in Figure 4 (and not according to the relative importance of these 10 risks).

3.1 RISK OF AI MISMATCH WITH BUSINESS NEED

Our common experiences have led us to highlight the risk of **non-alignment or inadequacy of the AI solution** to the identified use case. Quantifying the risk linked to the model implies doing so in relation to a **purpose** that is defined by the business and the use that it makes of the model. The **definition of the use and the limits** of the model are therefore essential. If this basis is poorly defined, the rest of the project will be impacted.

As presented in the previous section, the source of this risk lies in the combination of several factors, including:

- The **lack of business knowledge** about AI in general and the situations in which these techniques are relevant, the situations in which they are not, and the associated risks;
- The inaccurate **framing of the project** where the target variable is poorly defined in relation to the use case and the context;
- The **definition or choice of performance indicators** that do not correctly represent the performance of this model, particularly from a business point of view.

In this context, it should be noted that the media exposure of artificial intelligence tends to exacerbate this risk by encouraging people with little knowledge of artificial intelligence to get involved in projects related to artificial intelligence, without having the ability (i) to fully understand what it can bring them in general, (ii) to evaluate the difficulty of the tasks and (iii) to assess the relevance of the proposed solutions.

One of the most direct ways of reducing this risk lies in investing in **disseminating knowledge of AI and the associated risks** within the company, whether global acculturation (training in the main issues of AI and the system envisaged, understanding of the vocabulary adopted, etc.) or targeted towards the Model Developers to make them understand the issues of the businesses (customers of the AI development team).

Beyond the **awareness-raising aspect**, another measure concerns the management of AI projects, for example by setting up **transverse governance** that validates the **relevance of AI use cases** at the time of initiation once the scoping is complete. Finally, a last measure, albeit late in the development cycle of the use case, lies in the **quality control of the documentation** describing the needs, the objectives and the system chosen to meet them.

It should be noted that this risk is also reduced by the existence **of in-house AI teams**, which allow easier access to this expertise.

3.2 LACK OF RISK MANAGEMENT FOR THE USE OF PROTECTED AND/OR SENSITIVE DATA IN AI LEARNING

AI algorithms need a large amount of data for their training. The **collection of these data** is an important step prior to modelling.

Rules and controls need to be enacted and implemented to govern how this **data is collected and used** in the construction of an artificial intelligence system.

According to the various existing regulations, **data must be classified** as public, personal, sensitive, etc. The use of this data, apart from public data, in an AI algorithm must be **supervised** because the use (as well as the collection) of certain personal data (e.g., ethnic origin or sexual orientation) is forbidden. Their use could create discrimination and lead to reputational risk for the company.

This control of protected and/or personal data requires first of all external processes for **managing access rights to these data**. Protection measures must be put in place to guarantee confidentiality, access, and use. **Data masking solutions** can be envisaged if the use of certain data (anonymised or aggregated) is necessary, thus reducing the risk of identification of individuals and the disclosure of protected and/or personal data.

In order to meet these different needs at the time of data collection, particular attention, through a **risk management process and/or tool**, should be paid to the data that are necessary for the model.

In addition to the security and accessibility of this data, a certain discipline must be followed in order to limit the number of data used during the learning of the AI. Given the diversity and quantity of data likely to be used in the models, a first risk control measure is to ensure that the whole system benefits from **appropriate governance and management**. Institutions can rely on compliance mechanisms with existing regulations in this area, such as the general framework of the BCBS239 and the GDPR. In addition to this control system, it is possible to check how it is adapted to the volume, the sensitivity of the data and the diversity of sources in relation to the standard system. These checks can be based on interviews and reviews of operational controls (objectives and results). It should be confirmed that the collection system is sized for the atypical volume of AI: loading test, analysis of production incidents, review of logs, and control of quality rates of the most sensitive data.

The proliferation of input data to the model does not necessarily guarantee better performance, as the model becomes more complex to understand, not to mention the need for resources to train it. The **need-to-know principle** may be applied: only the data necessary for the purpose(s) should be accessible and used. Model Developers also need to be made aware of the issues surrounding the use of personal data in AI algorithms.

Although the rules defined by the various regulations must be put in place for "standard" data processing, the sheer volume of data processed in the creation of an AI algorithm can make it difficult to implement data controls that ensure full compliance with these regulations. Indeed, the controls and analysis of the data or its content cannot necessarily be carried out at the level of individual data. It is therefore important to put in place **controls that provide a global view** (identification and use) of the personal data used by the AI algorithm. These controls must themselves be capable of handling a large volume of data (input data, or processing logs). These controls must also include analyses of the quality of personal data, as their use, in addition to their regulatory framework, may generate unwanted effects (e.g., bias). These controls, if automated, should also be complemented by **regular reviews** to verify the relevance of the controls and their results.

3.3 Lack of identification of bias (direct or indirect) and creation or amplification related to the use of one or more input data in AI learning

When creating an AI algorithm, it is essential to understand the potential biases that may exist in the data used for training, and then be amplified by the modelling. The purpose of an AI algorithm is to reproduce a reasoning and/or prediction from data. This data comes from the real world, is sometimes, if not often, incomplete, and usually represents only a subset of a population from which the data was collected. By relying on this data, an **AI algorithm will inevitably tend to reproduce these biases.**

The first step is therefore to **identify them**, not only in the input data, but also to reduce biases that might be introduced in the modelling process itself. Indeed, modelling, which is necessarily based on assumptions and approximations, can influence (by reducing or amplifying) the discriminatory nature of the data.

It is essential to include the **identification of bias in the process of building an AI solution**. The identification of a bias will help to understand whether it will have an impact on the prediction and the objective sought. Identification can be done through the analysis of performance metrics on subgroups. For a classification task, there are different metrics for measuring bias in the model output. Among the most common, demographic parity ensures that, in the case of a granting model, the acceptance rate of the model for two sub-populations should be the same. Equal opportunity metric represents the same idea though conditioned on the population not failing. For example, the acceptance rate of the model should be the same for women and men who have paid back their loans well. It is often impossible to comply with all measures of fairness at the same time, as they may be incompatible. It is not always clear how much variation is allowed in the values of these metrics between two populations. One approach is to use the **disparate impact**. It corresponds to the ratio of a bias metric (e.g., demographic parity) between the protected and the preferred population. The 80% threshold or 4/5 law is commonly used, although this does not make it a regulatory threshold for AI models (except in the USA).

The second step is to **identify the source of the bias**, i.e., which variables are responsible for or may explain the difference in behaviour of the AI system. Comparing the distribution of variables in the model conditional on membership of a privileged or unprivileged population is an effective method for characterising and understanding the source of the bias.

The last step is to deal with the bias by **studying the possibilities of remediation** that will have a greater or lesser impact on the performance of the model. The objective is therefore to make the sensitive variables and the model outputs independent. There are **three types of approach**, always to be handled with care, so as not to risk creating other biases:

- **Pre-processing**: these methods aim to resolve the bias upstream, by modifying the training data via the deletion of variables, the addition of data or the weighting of observations. The deletion of a sensitive variable may be sufficient to correct the bias due to the correlations of the sensitive variable with other explanatory variables (association bias). The re-weighting technique reweighs the samples by group and label to aim for independence of the sensitive variable versus the label or class;
- **In-processing**: these methods aim to resolve the bias by modifying the training of the AI model, for example by adding a regularisation term in the cost function. It is often key to choose an objective (e.g., equal opportunity) as they are rarely all satisfied. Adversarial debiasing teaches an estimator to correctly predict the target label while minimising the ability of an adversary to predict the protected variable;
- **Post-processing**: these methods aim to resolve the bias downstream, at the output of the model. For example, the Reject Option Classification method corrects uncertain predictions according to the group of membership (protected or privileged). An alternative would be to calibrate a decision threshold per population.

The Big Data context implies that it is more complex to identify and deal with biases for multiple subgroups, i.e., for a large combination of sensitive variables. Eliminating multiple biases is virtually impossible. It is therefore a question of ensuring that data management and governance systems are put in place by the data office and compliance for the actors in the first line of defence (e.g., information and characteristics of the data, awareness of modellers, training and toolboxes to identify, measure and even compensate for a bias).

Some discipline in identifying bias should also be required at the modelling level, building on the techniques that we have briefly described in this section. This can be accompanied first of all by measures to make Model Developers aware of this problem and give them access to tools or libraries to facilitate this task. Tools exist and are offered by IBM (AI Fairness 360), Microsoft (Fairlearn) or Google (What-If tool). A wider awareness of the Model Owners / businesses is desirable to help them ask the right questions to the Model Developers during the discussions around modelling. However, an exhaustive search for all possible biases and their combinations remains a utopia.

Of course, a second check is part of the independent review of the model when the model is based on personal data (usually done by the second line of defence).

Finally, the third line of defence generally carries out reviews of the control mechanisms put in place by the other two lines of defence in order to ensure that they are working properly and that they comply with existing regulations.

The above considerations are applicable in the case where the model is developed end-to-end by the institution. The reuse of a pre-trained model increases the risks because detailed information is rarely available on the data used for the construction of the model or on the way the model was constructed (algorithm, hyperparameters, etc.). In view of the issues raised above and the forthcoming European regulation (AI Act), an AI solution provider should make available information about the data used and provide bias analysis results to ensure a minimum of transparency and results that will not create or amplify pre-existing biases.

3.4 MISUNDERSTANDING OR MISINTERPRETATION OF THE DATA USED IN THE MODELLING

The specific risk that we seek to highlight here relates to the **misunderstanding or misinterpretation of** the data used in the modelling.

This risk is not specific to AI, but it is more important because the volumes of data that can be used for modelling mean that there is a greater chance that some of the data used will not be correctly understood. In other words, the sheer volume of data input to the models limits the intrinsic analytical capacity as to its nature.

Furthermore, our current capacity to monitor data in existing infrastructures (source, transformation and normalisation) remains limited, particularly because of the disparate ways in which information systems evolve and in often heterogeneous environments. The correct interpretation of information is often subject to the availability of human experts with a working knowledge of that information.

This risk is particularly important because it is difficult to identify *a posteriori*. Indeed, the complexity of the models, or even their poor interpretability, does not always allow the identification of characteristics that may have a significant impact (quality, representativeness, measurement error, etc.).

In order to prevent these misunderstandings several control measures and good practices should be considered:

- **Meetings with the business** to understand the use case and allow the modelling team to check the data provided as input in terms of volume and distribution;
- **Building a comprehensive dictionary of data** used for learning, inference and monitoring (including filters);
- **Meetings with IT** to understand IT architecture, sources, data formats/units in the databases;
- The **independent model review process** also ensures that an independent view is taken of the variables used as inputs to the model.

3.5 RISKS ASSOCIATED WITH ONLINE LEARNING

Some systems that embed Machine Learning models incorporate a continuous automatic relearning mechanism, to allow the model to best adapt to changing queries and thus learn continuously rather than, for example, on a set frequency. This configuration is particularly relevant in use cases that can change rapidly, such as fraud detection. However, it is important to ensure that the configuration of continuous machine relearning is well thought out and that the **following risks are controlled**:

- **Risk of rapid divergence** of the model to be followed;
- Increased risk of model manipulation (cyber security);
- **Increased selection bias**: the model directly influences the selection of its own future training data.

It is necessary that a **control mechanism** exists for this continuous relearning and that it allows not only the **measurement of the risk** but also its **management**, i.e., for example the ability to switch to another potentially less efficient but more robust model if necessary. A potentially costly but effective control measure in relation to the problem mentioned consists of setting up a specific and independent (manual) sampling strategy, which makes it possible to measure

the performance of the model. For example, in the case of fraud, a subset of the transactions can be systematically reviewed by a human.

In reviewing the model, it is also important to estimate the extent to which the new training data over time corresponds to its own prediction (i.e. feedback loop).

Alternative methods can be used to monitor and compare the performance of a model over time, such as **creating one or more challenger models** (i.e., different solutions and/or implementations trying to solve the same problem) or **using an unsupervised model** to compare it to a supervised model.

3.6 DEFICIT OF INTERPRETABILITY / EXPLAINABILITY OF AI SYSTEMS

In a linear regression, it is easy to determine the weight of each variable and the positive or negative direction of impact from the coefficients. Similarly, the branch structure of a decision tree illustrates the sequence of rules that lead to the predictions. An AI model is often considered a **black box** because it is very difficult or impossible for a human being to predict the decision of the model on a new data element/point/instance. Thus this lack of intrinsic transparency can limit the ability of the business to validate whether an AI model makes business sense in terms of the variables selected and the influence of these variables on decisions.

The method of explainability must also be adapted to the purpose and the target audience. **Four levels of explainability** were proposed by the ACPR in June 2020[25] and a Tech Sprint on the explainability of AI algorithms was conducted in the summer of 2021. Among the results, the principles of intelligibility (i.e., trade-off between fidelity and sobriety of the explanation, conciseness of the explanation and trying to reconcile local and global explanations) and interactivity (i.e., adapting the explanation to the recipient's objectives) have been identified.

For a given use case, the data scientists (Model Developers) and the business (Model Owner) must jointly assess the degree of interpretability required, whether for model validation or transparency needs for the users of the AI system. Depending on the assumptions made about the relationship between the explanatory variables and the target variable, the degree of interpretability will change. Models based on linear and monotonic relationships or decision trees have a high degree of interpretability. On the contrary, models based on non-linear and non-monotonic relationships will have a very low degree of interpretability.

The main idea behind explainability techniques (XAI[26]) is to **provide a number of metrics** (e.g., importance of variables) or elements (e.g., rule extraction) that will allow modellers and the business to better understand the decisions of a model. Research in XAI is very active, and certainly not yet complete.

Explainability can be done at different levels. Global explainability allows us to understand the decisions of the model on average, through the classification of the importance of the variables (Feature importance), the form of the relationship between the explanatory variables and the target variable (Partial Dependence Plot), the interactions (or synergies) between the explanatory variables or the extraction of global rules (Anchors), via a simple substitution model (Global surrogate model). Local explainability allows us to understand decisions on a particular

[25] https://acpr.banque-france.fr/gouvernance-des-algorithmes-dintelligence-artificielle-dans-le-secteur-financier
[26] eXplainable AI : explicable AI

observation or a subset of observations. **Shapley decomposition and counterfactual analyses** are useful tools for local explainability.

Explainability can be either specific to a type of modelling (e.g., DeepLift for neural networks) or agnostic. The so-called agnostic explainability methods consist in separating the explanations from the model and are therefore independent of the type of methodology used. Depending on the explainability methods chosen (for example, LIME and SHAP), the explanation may be different or even inconsistent. It is therefore essential to ensure that the explanation can be trusted, by measuring its reliability, and to be aware of the assumptions and limitations of the explainability methods used.

To help Model Developers become more aware of the subject, it is recommended that tools be provided to **facilitate the construction of explanations**. The Shapash library[27], for example, allows the relevance of an explanation to be evaluated by calculating three metrics:

- **Local stability**, by comparing the explanation resulting from different explainability methods on similar instances;
- **Consistency**, by checking whether the explanations from different explainability methods are similar on average;
- **Compactness**, by analysing whether some of the variables are sufficient to explain the decisions of the model.

Theoretical documentation of the design, including the quantification of the influence of variables on predictions, providing documentation understandable by the business, and frequent interactions between data scientists and the business are essential to ensure transparency in the construction of an AI system.

3.7 DEPLOYMENT OF AN INSUFFICIENTLY STANDARDISED, SECURE AND CONTROLLED MODEL

After the model has been designed and implemented, it is **deployed in the information system** so that it can be accessed by users.

This deployment must be carried out according to the same deployment process as a "standard" application in the information system, i.e., follow the validation and control processes relating to the deployment of applications. Indeed, integration and validation tests are always necessary in order to ensure that the model functions correctly in the production environment, which sometimes differs from the development environment. These checks are necessary to detect possible effects with other existing models (e.g., in the case of mutually dependent models), incompatibilities (e.g. unvalidated versions of libraries not available in production).

Application deployment procedures and standards should support the deployment of AI models to ensure that a model does not lead to unidentified risks, e.g., related to cybersecurity, lack of or overuse of IT resources. Some modelling platforms now facilitate the deployment of a model, for example through the deployment of an API exposed on the same platform. In certain cases of preliminary testing, POC or by derogation, such a deployment may be accepted by the company's infrastructure and security teams, but this should either be formally and safely included in the company's IT procedures or remain an exception.

[27] https://github.com/MAIF/shapash

It should be noted that a **review of existing IT processes** may be necessary in order to integrate the specificities of AI and give access to model developers to a process adapted to their constraints. The sometimes-innovative characteristics of an AI model should not be an acceptable reason to avoid applying existing IT rules. It may be necessary to raise the **awareness of the teams** that will be using the modelling platforms in order to support them in the deployment of a new model. In this context, it is necessary to recommend the **implementation of MLOps processes** that allow for a stronger integration of business teams, AI teams and IT teams.

3.8 Lack of KPIs and/or lack of a monitoring process

Once the model is deployed, it is essential to monitor its relevance over time. If there is no **governance or relevant monitoring metrics**, a performance deviation may go unnoticed.

First of all, it is key to **define governance for monitoring models in production**. Indeed, monitoring metrics without a specified role expose a lack of reactivity in the actions to be taken. Thus, it is important to define and document responsibilities as well as the associated metrics, thresholds and actions. For example, if the performance of the model falls below a certain predefined level, the person responsible for monitoring will have to escalate in order to decide, for example, to re-train the model. In order for the actions to be carried out in a timely manner, the monitoring frequency must be aligned with the materiality of the use case. The granularity of the observations should be carefully defined: if the level of observation is too global, regional biases may go unnoticed. On the other hand, if the level is too granular, the comparison over time will be complex. Furthermore, the depth of the monitoring periods must be relevant in order to be able to compare the metrics to a reference period (such as the model development period). Broader periods can be monitored on an *ad hoc* basis, for example as part of a thorough annual back-testing exercise. Finally, the results of the monitoring should be discussed regularly between the various stakeholders, and in particular between the modellers and experts in the field.

Secondly, the **monitoring infrastructure** must be sufficiently **robust** to avoid any loss of information. The data used must reflect production, particularly in terms of completeness and granularity. Furthermore, the monitoring perimeter must be aligned with the application perimeter.

Thirdly, the **choice of metrics and associated breakpoints** is equally important. These must be aligned with the risk appetite while taking into account the model's own limitations. Different levels of alert and associated action are recommended. In addition, monitoring over time with comparison to the reference period is a key element for early detection of emerging deviations.

Global production metrics are used **to monitor risk exposure**, e.g., a strongly increasing volume of credit granting applications. In a Big Data context, it is important to control the availability and quality of data sources. If the number of variables in the model is high, specific quality checks can be limited to the most important features.

To avoid inconsistent model predictions, it is important to **monitor for outliers or missing data**. Indeed, some AI models or implementation packages are less robust to anomalies. In agreement with the governance, corrective actions can thus be decided.

The Big Data context also requires the average inference times to be controlled to avoid any latency problems due to a sharp increase in production volumes. Furthermore, as the model has been trained on a representative population, it is advisable to monitor the stability of the

data distributions. Indeed, these could evolve strongly towards areas where the error rate of the model is higher.

Further, the variables in the model must retain **their discriminatory power over time**. Thus, a variable that would no longer have a significant effect would increase the technical debt. The performance of the model must be measured using the metrics used in the modelling.

In addition, comparison with benchmarks based on alternative modelling can detect inconsistent predictions.

If a human-in-the-loop framework is applied, it is recommended to monitor the **rates of overrides** (exceptions) and their **justification**. This will help to understand certain weaknesses in the model and to make certain adjustments, e.g., recalibration of a threshold or addition of new explanatory variables.

If the model is applied to individuals, it is also recommended that **bias assessment metrics are monitored over time** to ensure that they remain within acceptable ranges.

Finally, it is necessary to understand whether the **factors influencing the decision** are always the same over time. This can be achieved, for example, by aggregating the Shapley values of local instances and comparing the results to the baseline period.

3.9 RISK OF BUSINESS TRANSFER

Just as it is important at the beginning of a project to ensure that the artificial intelligence approach is appropriate for the business needs, it is also important to ensure that when the model is transferred to the business, the results and explanations of the model are relevant and understood.

Risk control at the business transfer stage can be **multi-faceted**. To limit and control the risks inherent in this downstream phase of the use of AI, we have identified the following items:

- **Reviewing the change management plan** during the operational insertion of the model into the business process. This can be assessed directly through interviews and testing of actual usage. An indirect assessment can be made by analysing business incidents and their management (for models that already have a sufficient history of use);
- **Effectiveness of user training** assessed by interview or questionnaire to users, verification that users have access to the information necessary for proper use (existence of a user guide), analysis of the relevance of explanations, analysis of the consistency between user explanations and the reality of the theoretical model;
- **Existence of a risk mapping** and an adapted control plan;
- **Control of the actual use of the model results** and the relevance of model overrides (where human judgement is used to amend the outputs);
- **Existence of a monitoring system**. Where the model output is embedded in a software package, control is via authorisations. When the model produces more widely accessible data, the system can be included in the system for data governance.

Finally, the gaps between the planned deployment plan and the actual implementation can be analysed.

3.10 POORLY DEFINED GOVERNANCE: ROLES AND RESPONSIBILITIES

When transferring the use case to the business, it is important to properly **define the governance for the management of the AI system** over its lifecycle and to ensure that roles and responsibilities are clearly defined. Indeed, going into production does not mark the end of the work. Continuous monitoring **adapted to the materiality of the use case** must be put in place.

Model Risk Management systems generally define **key roles**, including that of the Model Owner, who must ensure the relevance of the AI system over time through a monitoring system.

Model change policies define the **typology of changes** (e.g., change of a parameter), their materiality and the need for an independent third party review (e.g. second line of defence). Defining this type of policy before going into production allows for smoother interactions between the various stakeholders and mitigates the risks associated with updates.

In the process of defining roles and responsibilities, the **control points** to be considered are, for example:

- **Rules for approving model recalibrations** and **informing users;**
- Clearly **identified business contacts** in the event of a malfunction, ability for first-level support to turn to the model designers in the event of an unexpected value;
- **Rules for updating user training and involvement of the model designers** for each modification. Control of the level of training and its updating;
- The **role of the process owner**: can be assessed by checking the correct positioning of the AI brick with regard to the process, particularly in terms of respecting the timetable for tasks if the brick tools a production process for example;
- The **processes to be followed** and the people to be involved when new versions of the libraries are put into production (and especially when old versions are withdrawn) in order to ensure the **compatibility** of the new versions and, if necessary, to change the models;

Finally, the **involvement and support of management** at the time of transfer is essential to ensure that roles and responsibilities are taken seriously. It is therefore necessary to ensure that management clearly communicates the main issues related to the operational insertion of the model.

4 CONCLUSION

Artificial Intelligence is a significant lever for **improving banking processes**, both in terms of creating value for customers and improving operational efficiency. However, it brings new risks that need to be acknowledged in order to properly manage and control them, and thus to take full advantage of the opportunities offered by the use of AI. The risk management and compliance processes in place in banks already cover many of the risks identified. The need to manage AI-induced risks is therefore not a **disruptive change** for banks or for certain so-called "critical" sectors (nuclear, aeronautics, autonomous vehicle, health).

On the other hand, other sectors that are less familiar with this type of system will have to **set up their own risk control framework**: the analyses presented here will hopefully provide them with food for thought in order to become more familiar with them and adapt them to their context.

The new risks thus created or increased by AI are notably related to the **use of data** (through Machine Learning), the **transformation and change management of business processes** that must be modified to incorporate AI models, **IT processes** when AI models are put into production, as well as **cybersecurity risks**, which are increased by the use of AI techniques that make attacks more effective, or even by new AI techniques ("adversarial" attacks).

In order to remedy this, certain actions must be taken, such as

- Establishing strong, high-level and cross-functional **governance**;
- Adapting and standardising **internal processes**;
- **Raising the awareness of all employees**, from the members of the Executive Committee to the operational staff;
- Promote the **proximity of the business, data science and IT teams**;
- Establish a **culture of risk control** to obtain an acceptable level of trust in AI from both internal users and external customers.

Today, the principles of risk control are largely formalised in banks, with an organisation in three lines of defence and a clear distribution of roles and responsibilities, constituting a solid foundation for the management of AI-induced risks. However, in practice, **the implementation of AI-related risk detection and mitigation analyses remains largely non-industrialised**, despite a constant search for process standardisation.

The risks identified in this white paper are **generic**, the approach adopted being based on an **inventory of AI-specific risks and a proposal of control methods**. As the assessment of their impact is specific to each AI model and each context of use, they are neither quantified in absolute terms nor prioritised. We thus deliberately refrain from providing a universal score grid allowing an overall assessment of the risk of a use case. On the contrary, the AI Act defines an *a priori* level of risk based on the uses, for a set of models beyond the scope of Machine Learning that we have set ourselves here. The regulatory obligations considered, for those falling under the "high risk" category in particular, aim to control the overall risks by means of ex ante control and documentation, in particular an audit trail that is particularly exhaustive and complex to implement. These requirements do not distinguish between risk factors and ad hoc measures to be considered. The control proposals presented in this white paper are therefore intended to complement the recommendations of the forthcoming regulations.

With the arrival of the **new AI regulatory texts** (and in particular the European Commission's AI Act[28]), companies, including financial institutions, will have to set up systematic and documented risk control processes. The cost of compliance could then drastically increase if the company does not quickly design a comprehensive risk control process that will accompany the production process of AI systems from end to end. Such a control process needs to be **standardised** and equipped with tools to make it more efficient and cost-effective in order to meet regulatory requirements.

By focusing on the three areas of governance, corporate culture and business expertise, companies will be able to prepare for the arrival of future regulatory texts and thus take full advantage of the benefits expected from the deployment of Artificial Intelligence.

[28] https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

5 GLOSSARY

BCBS239	*Basel Committee on Banking Supervision's standard number 239*: Principles for the aggregation of risk data and risk reporting
CNIL	Commission Nationale de l'Informatique et des Libertés (French Data Protection Authority)
KPI	Key Performance indicator
GPU	Graphics Processing Unit
Human-in-the-loop	Human intervention in the decision-making process
MLOps	A set of practices that aims to deploy and maintain Machine Learning models in production in a reliable and efficient manner.
POC	*Proof Of Concept*. Refers to an achievement whose purpose is to demonstrate the feasibility of a project.
GDPR	General Data Protection Regulation (EU GDPR), in French « Réglement général sur la protection des données » (RGPD)

6 SPECIAL THANKS

This White Paper is the result of a work started in 2021 by the Hub France IA within the framework of its Banking and Auditability Working Group.

We are especially grateful to the following people, who gave us their time and shared their experiences, both on the large group side and the start-up side.

Contributors

- **Léa Deleris**, Head of RISK Artificial Intelligence Research, BNP PARIBAS.
- **Jérôme Lebecq**, Data Science Coordinator, BNP PARIBAS.
- **Ludovic Mercier**, Inspection Générale, Directeur de pôle, LA BANQUE POSTALE.
- **Audrey Agesilas**, Supervisor – Model risk – Internal Audit, SOCIETE GENERALE.
- **Benjamin Bosch**, Manager - Model risk Management – Data Science, SOCIETE GENERALE.
- **Thomas Bonnier**, Model Risk Manager – Data Science, SOCIETE GENERALE.
- **Caroline Chopinaud**, Directrice Générale, HUB FRANCE IA.
- **Françoise Soulié-Fogelman**, Conseiller Scientifique, HUB FRANCE IA.

Reviewers

- **Nathalie Bouez**, Head of RISK Independent Review and Control, BNP PARIBAS.
- **Fabrice Le Chatelier**, Head of Data Science Office, BNP PARIBAS.
- **Michael Rabba**, Model Risk Senior Manager, BNP PARIBAS.
- **Rim Tehraoui**, formerly Group Chief Data Officer & Global ESG Risks Executive, BNP PARIBAS.
- **Pierre Contencin**, Responsable Validation des modèles, LA BANQUE POSTALE.
- **Emmanuel Jouffin**, Responsable du Département Veille Réglementaire Groupe, LA BANQUE POSTALE.
- **Fabien Monsallier**, Directeur innovation du Groupe, LA BANQUE POSTALE et Directeur Général, 115K.
- **Matthieu Olivier**, Chief Data Officer, LA BANQUE POSTALE.
- **Clémence Panet**, Chief Data Scientist, LA BANQUE POSTALE.
- **Julien Bohné**, Chief Data Scientist, SOCIETE GENERALE.
- **Anne-Cécile Krieg**, Deputy Head of Model Risk Management, SOCIETE GENERALE.
- **Julien Molez**, Group Innovation Data & AI Leader, SOCIETE GENERALE.
- **Eric Peter**, Head of Group model's audit, SOCIETE GENERALE.
- **Mélanie Arnould**, Chef des Opérations, HUB FRANCE IA.
- **Pierre Monget**, Chef de Projet, HUB FRANCE IA.
- **Andréa Arnaud,** Chef de Projet, HUB FRANCE IA.

Hub France IA

October 2022

BNP PARIBAS

LA BANQUE POSTALE

SOCIETE GENERALE

HUB FRANCE .IA