HUB FRANCE IA

A toolbox for managing risks of Artificial Intelligence systems

Banking & Auditability Working Group

October 2024

BNP PARIBAS

SOCIETE GENERALE

LA BANQUE POSTALE

TABLE OF CONTENTS

HUB FRANCE IA

1. Introduction

A toolbox for managing risks of Artificial Intelligence systems

October 2024

1. Introduction

The proliferation of uses of Artificial Intelligence (AI) is accompanied by the emergence of new risks. Most countries have therefore focused on providing a framework for the use of AI that should reduce its risks while reaping the benefits. In particular, the European Union has been working since April 2021 to produce regulations on AI through the *AI Act*[1]*(AIA)*. The legislative process has led to many amendments in the text proposed by the Council of the European Union[2]. The AIA was published in the Official Journal of the European Union on July 12, 2024, and came into effect on August 1st, 2024, while companies have 2 years to comply[3].

Compliance work requires considering the entire development process of an AI system, from the ideation, design, development, production phase to monitoring. Compliance costs are significant: in a survey[4] conducted by Hub France IA and its European partners in December 2022, more than 50% of the companies surveyed had indicated that they estimated the costs between €160k and €330k.

The Banking and Auditability Working Group of the Hub France IA, which brings together AI and audit experts from three major French banks, BNP Paribas, La Banque Postale and Société Générale, has been working for several years on managing AI risks. The working group shared its reflections and feedback on good practices in a White Paper[5] which presented the AI process and analyzed the risks.

With the arrival of the AIA and the rise of generative AI, we continued this work by trying to answer the following question: **how can risk management be equipped to reduce the costs of compliance** with the AI Act? The AI process could be supported throughout by a series of tools (documents, spreadsheets, software tools, etc.) which would ensure compliance validation as the process progresses and could then serve as proof for the supervisor (and in the specific case of financial

[1] European Commission. Laying down harmonized rules on Artificial Intelligence Act and amending certain union legislative acts. April 21, 2021. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75788

[2] Council of the European Union. Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Analysis of the final compromise text with a view to agreement. January 26, 2024. https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

[3] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonized rules on artificial intelligence. 13 June 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

[4] AppliedAI, Hub France IA et al. AI Act Impact Survey. December 12, 2022 https://www.hub-franceia.fr/wp-content/uploads/2022/12/AIAct-Impact-Survey_Report_Dec12.2022.pdf

[5] Hub France IA. Risk control of Artificial Intelligence systems. White Paper. October 19, 2022. https://www.hub-franceia.fr/wp-content/uploads/2022/10/22_10_19_Controle-des-risques-des-systemes-IA_PDF.pdf

institutions, proof for the internal control teams in charge of ensuring compliance with the AIA).

The approach adopted here was to describe the AI process as a whole, drawing on the White Paper[5] of 2022. At each stage, we have sought to identify a tool allowing to ensure compliance. As the AIA has just been released, it will be necessary to thoroughly analyze it before we can map the tools proposed here to the expectations of the AIA. This will be the subject of a forthcoming guide.

The work presented here is not intended to be exhaustive, which would have required a much longer document, but wishes to provide a **methodological framework and a set of good practices**. This work will hopefully be applied to other sectors of the economy, which, inspired by it, will be able to implement the processes adapted to their own context to better control the risks of the AI solutions they deploy or use and prepare for their compliance.

HUB FRANCE IA

2. The AI process

2. The AI process

2.1. Definition of AI

Artificial Intelligence is a "set of theories and techniques implemented with a view to creating machines capable of simulating human intelligence"[6]. It includes two main families: symbolic AI and digital AI, each of which has experienced periods of success and "winters", as shown in the diagram below representing the activity of these two families since the 1950s.

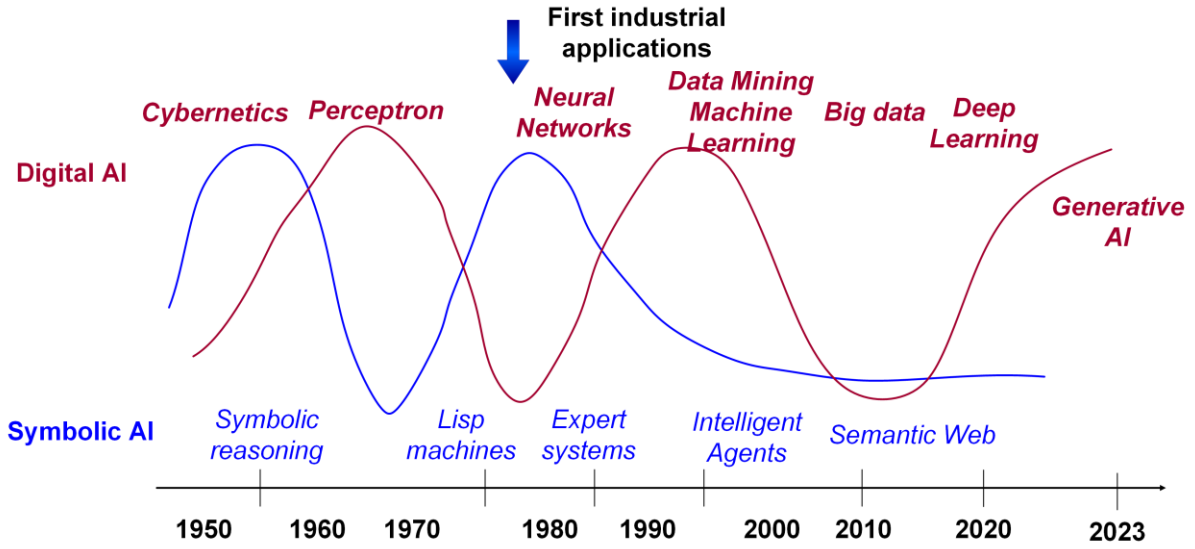

Figure 1 - Symbolic AI and Digital AI

In 2023, the launch of ChatGPT showcased the promises of Generative AI which we describe below. Symbolic AI is little used in the banking world, so we will focus on digital AI: the most widespread techniques today, and particularly in banking applications, are Machine Learning (or automatic learning) techniques, which include *Deep Learning* and *Generative AI* techniques. Subsequently, when we talk about AI, it will therefore be about *Machine Learning*, unless otherwise stated.

[6] https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257

2.2. AI in Banking

Artificial intelligence is increasingly supporting banking processes, and the trend has accelerated significantly in recent years. We refer the reader to our white paper for more details[5].

A significant proportion of AI use cases within banks are intended to **automate** internal **processes** in order to improve efficiency while reducing operational risk, to **improve the customer experience**, with the emergence of chatbots that assist customers in their operations, in customer knowledge (marketing) or in the design of new services for customers. Moreover, by improving the accuracy of algorithms, the use of AI contributes to **the mitigation of many risks, including operational risk, credit risk and compliance.**

Financial institutions are highly experienced in developing models, including controlling the modeling activity, because of the existing regulations in this field. However, it is important to pay attention when models are developed externally or in teams traditionally remote from modeling.

2.3. Machine Learning

The production of a *Machine Learning*-based solution is conducted in two stages:

- **Design** (or *Build*): based on an expressed need, the *data scientist* will collect the appropriate data to create a modeling dataset, then test various learning algorithms (very often from an open-source library). At the end of the learning process, we obtain an AI model - a program that can then be used: this program can be encoded in any computer language, the most common today being Python.
- Exploitation (**inference** or *run*): after the learning stage, the *data scientist* presents new data to the model built previously and obtains the result most probable for the data entered.

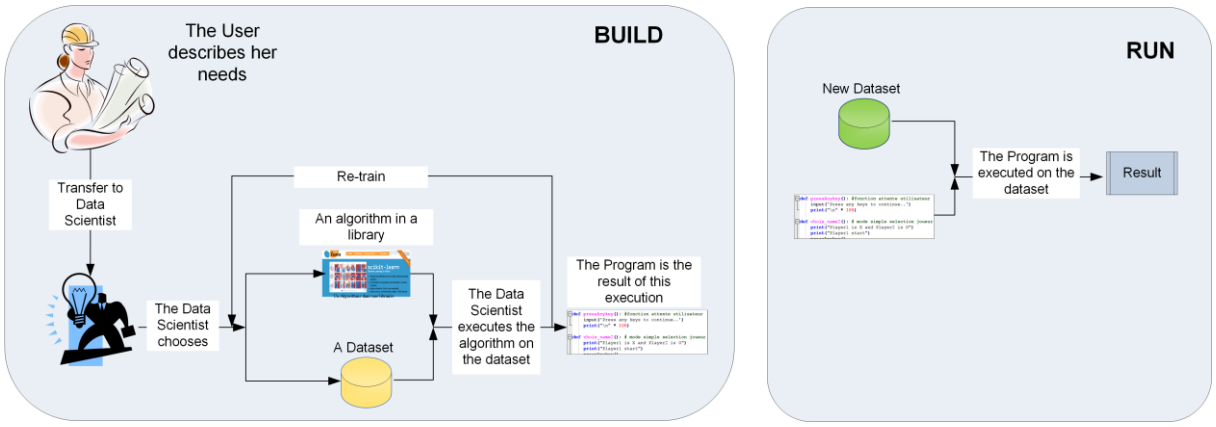

Figure 2 - The two stages of producing and using a *Machine Learning* model

It should be noted that three families of performance indicators are used in practice: **technical indicators** are used in learning to optimize the AI model, **business indicators** measure the business value generated by the use of the models, finally more **operational indicators** are also used, such as the duration of the calculation, the latency time, the number of variables and the complexity of the model, or even the cost of the variables if some are purchased.

2.4. Generative AI

Generative AI is a subset of Machine Learning, and more specifically Deep Learning, aimed at producing content, be it text, image, audio, or video, from input data (known as prompt), themselves text, image, audio, or video for example. A Generative AI model creates new content that is statistically consistent with the training data and the formulated prompt.

Generative AI models are generally trained on a very large set of data, requiring significant resources for their training. The architecture of the model (Transformer[7]) associated with the very large volume of data used during training allows for various uses for these models, without the latter having been trained specifically for these tasks.

Early solutions based on large language models, such as ChatGPT[8], can create text from input text instructions. These models are becoming increasingly multimodal,

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser. Attention is all you need. *Advances in neural information processing systems*. vol. 30, 2017. https://arxiv.org/pdf/1706.03762.pdf

[8] OpenAI. Introducing ChatGPT. Open AI. November 30, 2022. https://openai.com/blog/chatgpt

meaning that they can consider, both as input and output, data of several types, even combined such as image, audio, video, 3D, etc.

The generative capabilities of Generative AI models and their access continue to develop at high speed today, regularly paving the way for new uses.

The use of Generative AI in the banking sector is developing within the same framework as the use of Artificial Intelligence, mostly Machine Learning, i.e., in line with existing regulations in terms of internal control and risk management.

The banking sector, like many other business sectors, uses and manipulates a large volume of documents (text, table, image). Generative AI makes it easier to exploit large volumes of data, mainly textual, both for employees (search for information within a corpus of documents) and customers (answers more relevant and adapted to the context thus improving the customer experience).

The multi-modality of the models also allows unstructured data, such as voice or images, to be processed more industrially, while at the same time automating some very time-consuming tasks (e.g., customer protection compliance checks).

The use of Generative AI in the banking sector is progressing cautiously because, despite the possibilities offered by these new models, they are not free from defect, and their performance must be carefully assessed against the risks incurred before any widespread deployment of such models.

The main usage [9] categories therefore include: chatbots, augmented search systems (RAG: *Retrieval Augmented Generation*), systems for summarizing, support or content creation systems, systems for computer code and systems for reasoning on structured or unstructured data.

Below we explain some of the terms used in the literature.

It should be noted first of all that a Generative AI system can be produced in several ways as described in the figure below: either by directly using a commercial or *open-source* AI system, by refining (*fine tuning*) such a system, or by using an augmentation mechanism (*RAG*) exploiting a specific documentary base.

[9] Hub France IA. Uses of Generative AI. Volume 1 - LLM. January 2024. https://www.hub-franceia.fr/wp-content/uploads/2024/02/Livre-blanc_Les-usages-de-lia-generative-01.2024.pdf

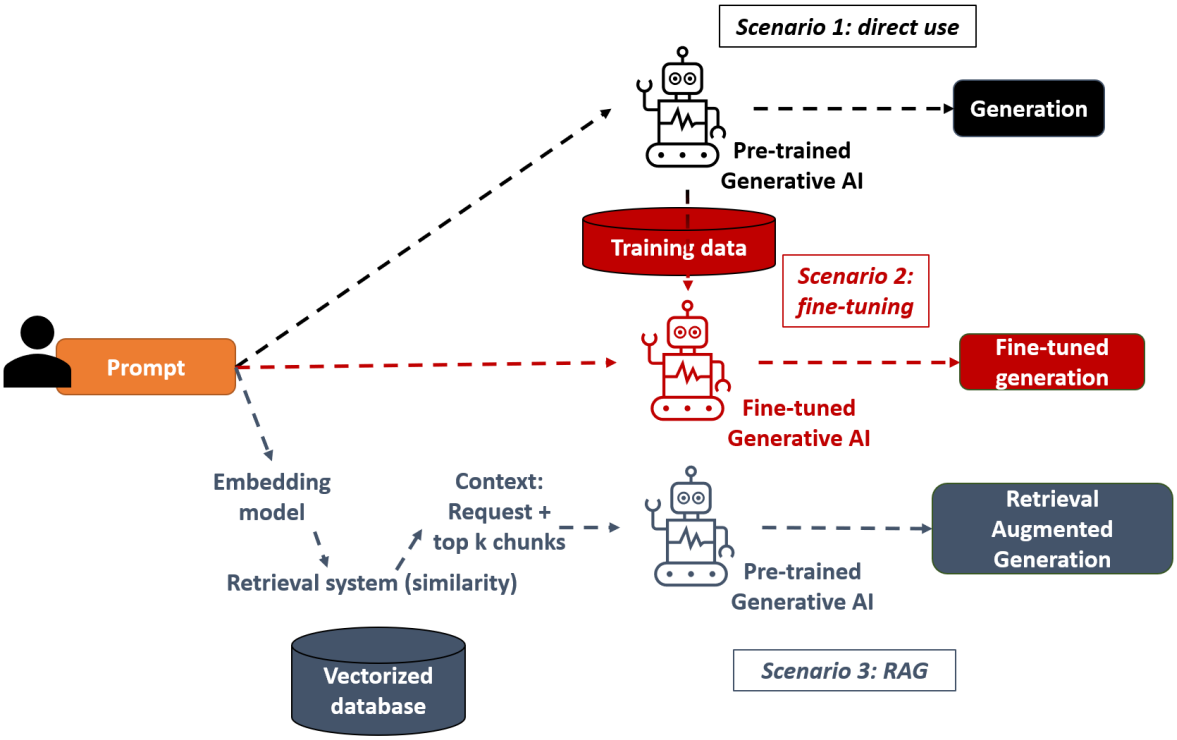

Figure 3 - Generative AI Application Scenarios

Prompts: The *prompt* is the natural language instruction or query provided to the Generative AI system for the purpose of obtaining a response. In general, the system prompt is used to give an instruction or context to the Generative AI (e.g., "you must answer in French"), while the *user prompt* allows to specify the question or query (e.g., "summarize this text"). Depending on the interface used (e.g., Application Programming Interface - API versus simple *chatbot*), it will be possible to specify for this *prompt* certain parameters such as the maximum number of tokens in the output or the temperature (which measures the degree of randomness in the generated text). In this context, *prompt engineering* is a strategy aiming at designing the most effective *prompts* to achieve the desired output. This strategy must adapt to each stage of the Generative AI lifecycle: *finetuning*, evaluation, or use of Generative AI by the business line. For example, when evaluating a Generative AI, it can be useful to generate a large volume of test prompt templates using a *Large Language Model* (LLM). It is also important to be clear in the instructions for the expected output of the task (e.g., 'Answer 'yes' or 'no') in order to automate the assessment of the LLM being tested.

Pre-trained Generative AI: a Generative AI is called "pre-trained" because it was built by training on very large volumes of data, thus allowing it to acquire general knowledge and the basis for generating relevant content. Large language models such as ChatGPT, Llama, etc. are pre-trained models. The integration in the prompt

of some examples of the task to be performed is called *few-shot learning*. This technique aims at increasing the pre-trained model performance on a specific task (e.g., classification, code generation). Specifically, the *prompt* will include one or more examples of user queries and assistant responses prior to the final Generative AI query. This technique is different from *fine-tuning* because the model parameters (weights) are not updated. It is just a matter of specifying the context by this *prompt engineering* technique. On the contrary, *zero-shot learning* consists in feeding Generative AI with a direct prompt, without an example of the task to be performed.

Embedding: An *embedding* is a vector representation of large dimension for an object such as a word, document, or image. These representations are learned during model training. They contribute to the calculation of attention in the Transformers models (at the basis of LLMs) and make it possible, for example, to know which words are semantically close. Pre-trained *embedding* models thus make it possible to vectorize objects in a relevant manner in order to carry out research tasks in particular (e.g. RAG). In the case of RAG, for example, the user's text query (his prompt) is transformed into a vector by such an *embedding* model. This vector is then compared with other *embeddings* stored in a vectorized database. Measures of similarity make it possible to extract and order the top k most relevant sections of the base (k = 1, 2, 3, 5, 10 for instance, k being a parameter chosen by the persons building the specific solution). With the concatenation between the k sections and the query, Generative AI can thus generate the most relevant response. This shows the essential role of *embeddings* in the case of RAG.

RAG (*Retrieval Augmented Generation*). The knowledge of a Generative AI system does not go beyond the scope of its training data. For the responses generated by the AI to be based on fresher and/or more specific data, it is necessary to give it access to a **knowledge base**. The construction of this base is carried out in several stages. First, the text data must be obtained to feed this database (e.g., PDF files). Next, the data should be divided into chunks of text with a given strategy (e.g., fixed number of *tokens*) which can overlap. The third step is to vectorize these chunks of text using a pre-trained *embedding* model. This vector representation helps the search process because chunks of similar text have similar representations. The next step is to create the vector database with an index to which the previous *embedding* vectors are stored. This data structure enables to quickly retrieve the relevant information. As described in the section on *embedding*, it is then possible to couple this *retrieval system* with a Generative AI to perform retrieval augmented

generation. The user sends a search query, which is vectorized with an *embedding* model. The retrieval system selects the top k *chunks* using semantic search methods based on dense or sparse vectors. Finally, AI responds to the user based on the context and using its generation capabilities. The context thus contains the query, and the top k most relevant *chunks* selected in the vector database. RAG therefore provides more precise answers on the area defined by the knowledge base.

Fine-tuning: *fine-tuning* a pre-trained Generative AI consists in continuing to train it on specific labeled data in order to improve its performance on a particular task or field. Unlike the "few-shot learning" or "zero-shot learning", the parameters (weights) of the pre-trained model are thus updated. For example, Generative AI can be specialized on a classification task related to a particular domain. Another example: a chatbot in the Bank can be refined by retraining on series of questions / answers labeled as "good answer" or "bad answer" by customer advisors. *Fine-tuning* thus requires a certain investment, the real added value of which must be assessed in comparison to an experienced prompt strategy. The first step in *fine-tuning* is to prepare the training and validation data. For example, it will be a series of prompts, each containing roles (system, user, assistant) and content (instruction, question, answer). The Generative AI model is then trained and evaluated on each respective data partition. If the performance achieved is not satisfactory, it is then desirable to continue training with targeted prompts, for which the model does not offer the expected results. Finally, the Generative AI model can be used on the targeted use case. *Fine-tuning* is expected to reduce hallucinations in the particular field considered.

LangChain: *LangChain*[10] is an *open-source* tool to streamline the development of applications around Generative AI. In particular, *LangChain* makes it possible to create pipelines involving one or more Generative AI models with a sequence of tasks such as text generation and then translation. Similarly, it is possible to create a *retrieval chain* to integrate *RAG* into the sequence.

Guardrails: These are protections which allow to control the inputs and outputs of a Generative AI in order to reduce the risks related to its use. They are particularly important in the context of an application used externally by customers (e.g., *chatbot*). Risks can be linked to various aspects, such as toxicity, bias, cyber risk,

[10] https://www.langchain.com/

data leakage, or hallucinations. Generative AIs have generally incorporated a well-defined ethics policy. However, it is sometimes possible to circumvent it by creating adversarial prompts. To reduce these risks, it is useful to give instructions on ethical behavior to the Generative AI via the *prompt* system or to use a control tool such as *LangChain Constitutional AI*[11]. In this latter case, a trade-off between latency and level of risk will have to be decided.

[11] https://python.langchain.com/docs/guides/productionization/safety/constitutional_chain/

2.5. The model production process

The process of producing a predictive AI model, described in the figure below, involves various stages, with the potential for iterative backtracking until you are satisfied with the result.

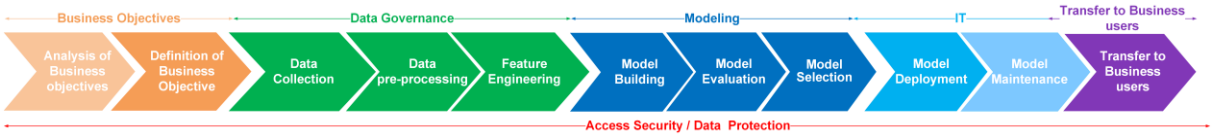

Figure 4 - Predictive AI model production process (linear representation)

The process of producing a Generative AI model is described below. However, it should be noted that the process will be slightly different depending on whether the model is taken directly "*out of the box*" (Copilot 365[12] for example), fine-tuned or by exploiting a RAG method. In addition, as the field evolves very rapidly, the process itself is likely to evolve:

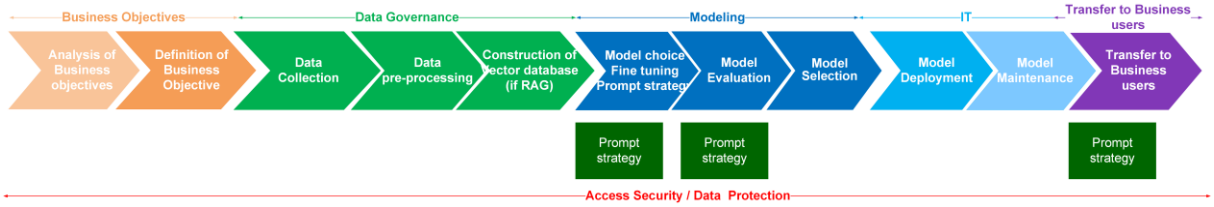

Figure 5 - Generative AI Model Production Process

In this diagram, we specifically highlight the stages in which the *prompting* strategy intervenes, whether it is used to model or is evaluated. The risk assessment of an AI solution will be done throughout the process, one needs to ensure that the right tools and the right players are positioned where they are needed, as we described in our white paper[5] on risk control.

[12] Colette Stallbaumer. Introducing Microsoft 365 Copilot—A whole new way to work. Microsoft. March 16, 2023. https://www.microsoft.com/en-us/microsoft-365/blog/2023/03/16/introducing-microsoft-365-copilot-a-whole-new-way-to-work/

HUB FRANCE IA

3. AI Act certification

A toolbox for managing risks of Artificial Intelligence systems

October 2024

3. AI Act certification

The regulation of artificial intelligence systems (AIS) as implemented by the Artificial Intelligence Act[13] (AI Act) is based on a risk-based approach. Risk is defined as "the combination of the probability of an occurrence of harm and the severity of that harm"[14]. The regulation then distinguishes four levels of risk:

- Unacceptable risks
- High-risk
- Other (with transparency requirements)
- Other (without transparency requirements)

This first classification applies to AIS falling within the scope of the text (i.e., meeting the definition of AI system in the AI Act). This paper focuses on predictive AI systems, which are a subset of AI systems that fall within the scope of the AI Act.

The regulation of so-called "**general-purpose**" AI models was added later and by a similar approach. These models are divided into two levels of risk within the AI Act: **General-purpose AI models with systemic risk** and the others.

Depending on the level of risk corresponding to the predictive AI system or the general-purpose AI model, **the applicable requirements and obligations differ**. These also evolve according to the position of the operator in the **value chain of AI**. The AI Act thus distinguishes the providers, deployers, authorized representatives, importers, and distributors of an AIS[15].

The two main roles associated with the implementation of an artificial intelligence project within a company are provider and deployer.

[13] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 13 June 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

[14] *Ibid,* Article 3§2

[15] It is important to note that an operator may cumulate or alternate between these statuses over time and during its activities.

Provider: "a **natural or legal person**, public authority, agency or other body that **develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market** or **puts the AI system into service under its own name or trademark**, whether for payment or free of charge"[16]

Deployer: "natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity"[17]

As part of this deliverable, we will therefore focus our developments on the obligations imposed on these two operators, particularly when designing and/or deploying a high-risk predictive AI system or a Generative AI.

3.1. Requirements and obligations for predictive AI systems

3.1.1. Classification of predictive AI systems

As mentioned above, predictive AI systems are divided into four categories within the AI Act (this table does not provide an exhaustive summary of all nuances existing in the regulation):

Risk category	Characteristics of AIS
Unacceptable Article 5	Uses subliminal techniques beyond a person's consciousness **Or** Exploits a person's vulnerabilities **Or** Uses biometric categorization systems to infer/deduce sensitive data about the individual **Or** Uses a social *scoring* technique **Or** Uses remote biometric identification in real time for law enforcement purposes

[16] *Ibid,* Article 3 § 3.
[17] *Ibid,* Article 3 § 4.

	Or Uses a technique to predict the risk of wrongdoing (Note: this list is not intended to be exhaustive and each of the above cases is subject to precise conditions)
High-risk Articles 6 to 15 Article 27 Listed exceptions Article 6(3a), (3b), (3c) and (3d)	Is referred to in Annex III
	Is covered by one of the EU harmonization legislations listed in Annex II of the AI Act **And** Must undergo a third-party conformity assessment before it is placed on the market. **And** Need to assess the fundamental rights impact of high-risk AI systems
Other (with transparency requirements) **Article 50**	Transparency requirements for providers and users of certain AI systems.
Other (without transparency requirements)	All AIS not covered by the above criteria.

For the high-risk category, exceptions are listed (recital 32a Article 6, paragraphs 2a and 2b):

If an AIS

Meets the criteria for categorizing high-risk AIS but does not pose a significant risk of harm to the health, safety, or fundamental rights of natural persons, including by not materially influencing the outcome of decision making.

The following AIS are subject to this exemption:

- Is intended to perform a narrow procedural task

And/or

- Is intended to improve the result of human activity

And/or

- Is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review

And/or

- Is intended for a preparatory task prior to human assessment

AI systems **profiling** natural persons are still considered to be high-risk.

3.1.2. Requirements and obligations for high-risk AI systems

The classification of an AI system as "high-risk" can thus result from two scenarios:

- AIS shall be used in one of the areas listed in Annex III because of the **harm it may cause and its severity on the health, safety, or fundamental rights of natural persons.** These areas are, for example, the management of certain critical infrastructure, the systems used to assess the creditworthiness of natural persons or to establish a credit rating, etc.
- The AIS is used as a safety component of a product or is itself such a product, covered by Union harmonization legislation (toys, maritime vehicles, etc. see Annex I to the AIA).

These high-risk AI systems will be subject to requirements and obligations to be placed on the European market. The AI Act first distinguishes between the requirements, i.e., the characteristics and prerequisites applicable to the AIS itself, and the obligations.

The table below summarizes, in a non-exhaustive way, the different requirements applicable to high-risk AI systems.

Category of requirement	Requirements Details
Risk management system Article 9	Establishment and implementation of a documented and up-to-date risk management system throughout the life cycle of the AIS.
Data and data governance Article 10	Training, validation, and testing data sets shall meet quality criteria. In addition, they must be subject to governance and data management requirements.
Technical documentation of the model Article 11 Article 72 Annex IV	The **technical documentation shall be drawn up in such a way as to show that the high-risk AIS meets the requirements**. It shall be drafted in a **clear and intelligible** way to be used for the conformity assessment of the AIS. It shall include at least: • General description of the AI system, • Detailed description of the elements of the AIS and its development process, • Detailed information on the monitoring, functioning and control of the AIS, • Information on the appropriateness of the performance metrics for the AI system concerned, • Description of the risk management system in place • List of harmonized standards (not published yet) applied in full or in part, • Copy of the EU declaration of conformity, • A detailed description of the system in place to evaluate the performance and monitor AIS in the post-market phase (Art. 72).
Record-keeping Article 12	Automatically records events throughout the system lifetime. Logging capabilities should record relevant events to **identify high-risk situations, facilitate post-market surveillance**, and **monitor the functioning of the AI system**.
Transparency and instructions for use Article 13	The AIS should be designed and developed in such a way as to promote a level of transparency that allows

	deployers to use the AIS appropriately and to interpret the results. It is accompanied by instructions for use for deployers containing: • Identity and contact details of the supplier, • High-risk AIS characteristics, capabilities, and performance limits, • Changes in the AIS and its predetermined performance by the supplier at the time of the initial conformity assessment, • Human oversight measures, • The necessary IT and hardware resources, the expected lifetime and the necessary maintenance and monitoring measures to ensure the proper functioning of the AIS, including software updates, • Where applicable, a description of the mechanisms included in the high-risk AI system that allows deployers to collect, store and interpret logs correctly.
Human oversight Article 14	The design and development of high-risk AI systems shall allow, in particular through appropriate human-machine interfaces, effective control by natural persons during their period of use. This human oversight aims to **prevent or minimize the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used.**
Accuracy, robustness, and cybersecurity Article 15	The design and development of high-risk AIS shall ensure an appropriate level of accuracy, robustness, and cybersecurity, and shall operate consistently in this respect throughout their life cycle.

Table 1 - Requirements for high-risk AI systems

These requirements must be put in place before (*ex-ante*) the placing on the market/putting into service of the high-risk AIS and remain throughout the life cycle of the AIS.

The Artificial Intelligence Regulation then details **the obligations for providers and deployers of these AI systems.** It is important to note that if the deployer

commercialize or substantially modifies a high-risk AI system already placed on the market or put into service in its own name or brand, the deployer is then considered as the provider of the AIS (Article 25).

The following table summarizes in a non-exhaustive way the different obligations applicable to high-risk AI system providers.

Obligations for high-risk AIS providers	Details of obligations
Compliance with requirements Article 16	The supplier shall ensure that the AIS complies with the requirements set out above.
Establishment of a quality management system Article 17	The supplier shall set up a quality management system to ensure compliance with the AI Act. It shall be documented in the form of written policies, procedures, and instructions. It shall include at least: • A strategy for regulatory compliance, • Systematic techniques, procedures, and actions to be used for the design and control of high-risk AIS and to verify their design, • Techniques, procedures, and actions to be used for the development, control, and quality assurance of the AIS, • The examination, testing and validation procedures to be carried out during the life cycle of the high-risk AIS and their frequency, • The technical specifications, standards and means to be used to ensure that the high-risk AIS complies with the requirements of the AI Act, • The risk management system put in place, • Details of the setting-up, implementation and maintenance of the post-market surveillance system, • Procedures for the report of a serious incident (Article 73), • Handling of communications with competent authorities,

	• The relevant systems and procedures for keeping documents and information, • Management of resources, • The accountability framework.
Documentation keeping Article 18	Documentation from Articles 11 and 17 of the AI Act, as well as documentation on change approvals and decisions issued by notified bodies and the declaration of compliance with the AI Act, shall be retained for 10 years after the placing on the market/entry into service of the high-risk AIS.
Automatically generated logs Article 19	The provider shall retain the automatically generated logs referred to in Article 12. These logs shall be kept for a period of at least 6 months appropriate to the destination of the high-risk AIS.
Corrective actions and duty of information Article 20	Where a high-risk AIS provider has reasons to consider that one of its AIS placed on the market/in service is non-compliant, it shall immediately take the necessary corrective actions and inform the deployers and other relevant actors thereof.
Cooperation with competent authorities Article 21	At a legitimate request of a competent authority, the supplier shall make available to that authority all the information and documents necessary to demonstrate the compliance of the high-risk AIS.

Table 2 - Requirements for high-risk AI system providers

Finally, Article 16 of the AI Act specifies **the procedure to be followed before a high-risk AI system is placed on the market/put into service**. Thus, to be placed on the market, these systems will have to undergo a **compliance assessment procedure** (Article 43), **a declaration of compliance** (Article 16), be **affixed the CE marking** on their packaging or documentation (Article 16) and **be registered in the EU database.**

As regards AI system deployers, these obligations are mainly organizational (Article 26). For example, the deployer of a high-risk AI system must, in a non-exhaustive manner:

- Take **appropriate technical and organizational measures to ensure that it uses the high-risk AIS in accordance with the instructions for use of the system**,
- Ensure that the persons responsible for the human oversight of the system have **the skills, training, and authority** necessary for their tasks.
- **Control the input data** and ensure that it is relevant and sufficiently representative for the purpose of the high-risk AIS,
- **Monitor the high-risk AIS** based on the system notice and **inform the supplier of any event that may present a risk,**
- Ensuring the maintenance of high-risk AIS logs,
- Inform workers' representatives and workers who will be subject to the use of the high-risk AIS.

3.2. Applicable regulations for Generative AI models

3.2.1. Regimes applicable to Generative AI models

The concept of a "general-purpose AI model" came late in the development of the AI Act.

> General-purpose AI model is "an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market."[18]

Under this definition, Generative AIs fall under general-purpose AI models. Generative AIs are therefore regulated by the provisions of the AI Act on general-purpose AI models but also by the provisions on AI systems generating or manipulating images, or audio and video content (Article 50).

[18] *Ibid,* Article 3 § 63.

3.2.2. Classification of general-purpose AI models

The provisions applicable to general-purpose AI models (Articles 51 et seq.) divide these models into two categories: general-purpose AI models with systemic risk and the others.

The notion of **systemic risk** is specific to these AI models and is understood as a risk of "having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain"[19].

A general-purpose AI model is classified as having systemic risk when it has "**high-impact capabilities**"[20]. This is assumed if the model meets one of the following conditions:

- 'The cumulative amount of computation used for its training, measured in floating point operations, is greater than 10^{25}' [21].
- It shall be the subject of a Commission decision, of its own motion or following a qualified alert by the Scientific Panel based on the criteria identified in Annex XIII, including, for example, the number of parameters of the model, the size or quality of its dataset or the number of registered end-users.

3.2.3. Obligations of providers of Generative AI models

For Generative AI, the obligations of their providers are similar to those of high-risk AI system suppliers.

For Generative AIs assimilated to general-purpose AI models (excluding open source), they consist in the following[22] elements:

- Draw up and keep up to date technical documentation of the model and information for AI system providers considering integrating the model into their AIS,
- Establish a policy to comply with European copyright law,

[19] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 13 june 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689.
[20] *Ibid*, Article 51(1a).
[21] *Ibid*. Article 51(2).
[22] See in particular Article 53 et seq.

- Make public a detailed summary of the content used to train the general-purpose AI model,
- Cooperate with the European Commission and the competent national authorities.

For **Generative AI systems with systemic risks**, there is also an obligation for suppliers to carry out model assessments to identify and mitigate risks and ensure an appropriate level of cyber security protection (Article 55).

Finally, **for all Generative AIs**, all **content generated by Generative AIs** (text, image, etc.) will have to be marked "machine-readable format and detectable as artificially generated or manipulated"[23].

3.3. Meeting the AI Act Requirement Call for an industrialization of the compliance process

The Artificial Intelligence Act provides for severe penalties for non-compliance with its requirements and obligations. These sanctions are intended to ensure the implementation of the Regulation and may take the form of warnings or fines. They are intended to dissuade operators subject to the AI Act from violating it.

For example:

- Failure to comply with the obligations incumbent on the supplier or deployer shall be punished by an administrative fine of up to **15 000 000 euro or 3% of the total worldwide annual turnover** achieved during the previous fiscal year of the undertaking.
- The provision of inaccurate, incomplete, or misleading information to notified bodies or national authorities in response to their request shall be punished by an administrative fine of **up to 7 500 000 euro or 1% of the total annual worldwide turnover** in the previous financial year.

For these examples, the highest figure is the one to be used for the calculation of the amount of the fine.

There is no doubt that, **to comply with the Artificial Intelligence Act, companies will benefit from industrializing the process**.

[23] *Ibid,* Rule 50.

HUB FRANCE IA

4. Operationalization of risk management

A toolbox for managing risks of Artificial Intelligence systems

October 2024

4. Operationalization of risk management

4.1. Methodology

Similar to the previous white paper[5] from our working group, we structure our analysis around the AI system production process from ideation to monitoring (figures 4 and 5) .At each stage, we seek to identify the tasks to be performed in order to establish the artefacts that would be necessary for a certification audit: these artefacts must be aligned with the obligations listed in the tables above, in "AI Act certification" chapter (Note that this document was drafted before the final publication of the *AI Act*, the delegated acts and the forthcoming harmonized standards). These elements can be of different kinds: a guide or procedure, a spreadsheet, a program / code, a score, a template / standard, a list or form to be filled in, a workflow tool, etc. They enable at each step to obtain a piece of evidence to support the certification audit: for example, a spreadsheet will allow AI systems to be scored and then triaged according to their expected risk level, and a procedure will then indicate how to select an AI system from a list of potential cases.

In the following paragraphs, we present these different elements **as we move through the successive stages of the AI system's production process**. We have naturally aligned with successive versions of the AI Act to define these elements. But the AI Act was not yet official when we published the French version of this white paper, so there could be some minor misalignment with the requirements of the AI Act: this work remains to be done, including a legal component (which we have not considered here) to achieve compliance with the regulations of the AI Act.

All these artefacts, together with the proposed methodology, constitute the necessary tools for the operationalization of risk management with a view to being compliant with the AI Act.

Ideally, software tools would be proposed on the market to implement and coordinate the production and storage of some or all of these artefacts. In a final paragraph, we present an analysis of a few commercial tools offering such solutions.

Before developing further, we present in Figure 6 a graphical summary of the artefacts that we identified as required for good risk management of AI systems. They are organized by type of artefacts and by stages in the production process of an AI system. Some of these elements structure the framework around risk

management (procedure, guide, templates, scores and sometimes code), others are deliverables specific to each use case, in particular documents, lists and inventories and finally proofs of workflow in decision-making processes. In the rest of the text, these elements appear in **red** letters.

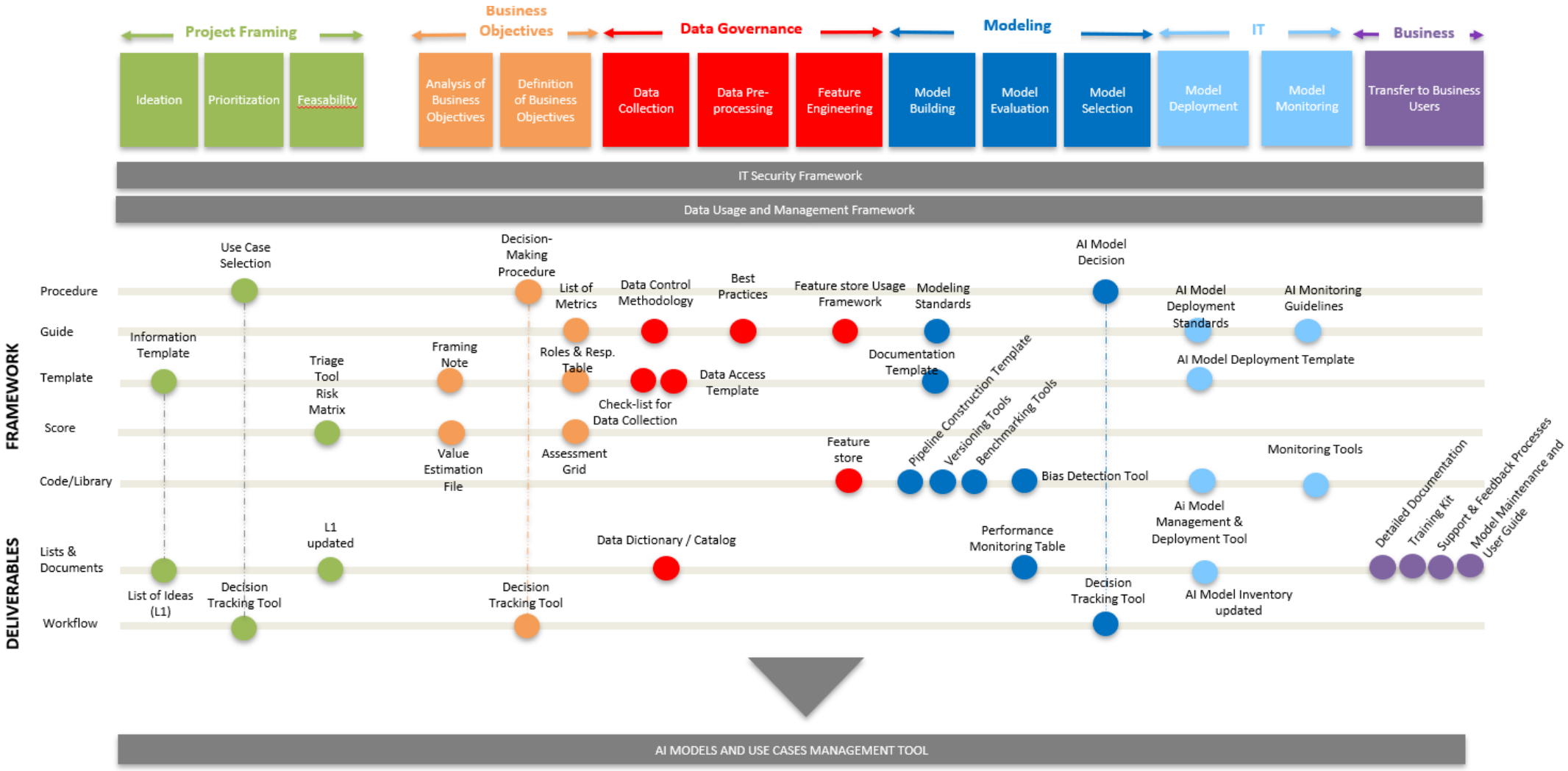

Figure 6 - Summary

4.2. Project Definition

4.2.1. Ideation

The process of ideation focuses on generating AI use case ideas. Schematically, it entails organizing workshops with one or more people belonging to a specific business area and to encourage, through a moderation by experts, the identification of pain points or opportunities that could be addressed by AI.

To support the ideation process, it is useful to set up a tool that stores all ideas of AI use cases in order to subsequently allow them to be prioritized. Indeed, without coordination or prioritization, investments to develop use cases will be allocated on an ad hoc basis and will hardly be aligned with the company's strategic objectives (whether financial or ethical).

INFORMATION TEMPLATE

In order to structure the collection of ideas, it is recommended to define a **Template** to be collected for each use case idea developed.

LIST OF IDEAS

At the end of the ideation process, **a list of ideas** described according to the defined template will be obtained. This list serves as a basis for the prioritization process.

4.2.2. Prioritization

The objective of the prioritization process is to identify which use case ideas will be selected for further development. It is therefore a matter of selecting and sequencing the investments in the development of AI.

USE CASES SELECTION PROCEDURE

To support this prioritization process, it is useful to define **a selection method**, reflecting a value framework that allows the contribution of each use case to be quantified, for example through a **procedure and/or a decision-making committee.**

DECISION TRACKING TOOL

Once decisions are made, both the elements of the process (date, location, contributors, etc.) and the outcome of the selection should be documented in a record keeping tool.

4.3. Feasibility

4.3.1. Risk analysis

Even at this early stage, it is important to look at a rudimentary analysis of the risks associated with the use cases proposed in order to identify: those that are not acceptable (often for ethical, legal and regulatory reasons); those for which development must be closely monitored in order to identify the sources of risks and their materiality as they arise and for which risk control solutions should be put in place at the design stage; and those that present a priori low risk, for which the risk profile must nevertheless be monitored but in a less sustained manner than for the others. Finally, it is necessary to determine beforehand whether there is a risk that the project will not be technically implementable (e.g., heavy dependence on external data, inappropriate frequency of data supply, hardware requirements).

This risk analysis may in particular affect the roadmap or the priorities that have been defined in the previous stages.

TRIAGE TOOL (SCORE)

In concrete terms, this involves defining a **simple triage tool (score)**, which can, for example, be based on what is proposed in the US regulations SR 11-7, i.e. looking globally at three axes (i) the impact of the use case, (ii) its complexity, (iii) the capacity to monitor performance when the system is in production. It can simply be a matter of defining for each of the metrics' axes which are then associated with a score, for example between 0 and 1, and which is aggregated over the three axes.

RISK MATRIX

Another example is to define a matrix that produces a qualitative level of risk (low/medium/high) according to the probability of occurrence of the risk and its severity.

Note that if the business comes with a specific / identified need without going through the ideation phase, it will nevertheless have to conduct a feasibility study.

4.4. Business objectives

4.4.1. Analysis of the Business requirements

We are now in the context of a specific project.

The starting point of a project is to accurately frame the business needs. To this end, it is useful to structure this discussion around **a list of questions** that includes the business requirements, in particular, the expected benefits, the availability and quality of data, the ability to deploy the solution both from a technical point of view and especially from an organizational point of view (impact on processes, usability, acceptability, social issues). Naturally, this list of questions is more exhaustive than the one defined in the **Template** associated with the ideation phase.

OUTLINE

The result of this step is therefore **an outline** of business requirements that could be consulted or updated throughout the project. It is therefore also appropriate to continuously monitor and assess changes to this note.

One important but complex point at this step is to estimate the value added by the project. This estimate is based on an analysis of the existing solution that must be done by the client (who knows the current process) and then on the target solution that must be done jointly by the development team and the client's team.

ADDED VALUE ESTIMATION FILE

To ensure comparability between projects, it is suggested to structure this value estimation by **a procedure** and/or **a tool** (which will probably be a spreadsheet at first). It is also useful to make available various charts (such as the hourly rate per country, time spent on certain common tasks, etc.) in order to standardize the values of the parameters used in the estimation.

4.4.2. GO / NO GO Decision

ASSESSMENT GRID (SCORE)

Following the scoping phase, an official decision of whether to launch the project is required. This decision can be based on **an assessment grid** (score) that combines previous analyzes (initial requirements, risks, benefits, data availability, technical

feasibility, impact on processes, deployment requirements, preliminary risk management model) with considerations of availability of technical teams.

ROLES AND RESPONSIBILITIES TABLE

This assessment grid also provides a clear specification of the roles and responsibilities, which are ideally included in a **predefined table of roles and responsibilities**. Beyond the model developer and model owner roles, it is useful to clearly specify who is responsible of obtaining the data, managing the change, or interacting with IT teams of impacted systems.

DECISION-MAKING PROCEDURE

Similarly to a prioritization process, it is useful to define **a decision mode** through a **procedure** for instance, to support the decision-making process**,** and once the decision has been taken to document both the elements of the process (date, place, contributors, etc.) and the result in a record keeping tool.

4.4.3. Specifying Business Objectives

LIST OF METRICS

If the project is selected, then the business objectives specification continues by selecting the metrics to be followed during the project. To simplify and standardize this process within the organization, **two standard lists of metrics** can be provided:

- A first one that includes **business-oriented metrics** (business performance, volumes of data to be processed, constraints on exploitability, latency).
- A second one that lists **the data science-oriented metrics** (performance metrics, constraints on complexity and explainability).

The purpose of this step is therefore to produce **the completed lists of business and data science metrics** that must also be recorded and monitored over time.

In the specific case of Generative AI, in order to evaluate and compare different LLMs on similar *Massive Multi-task Language Understanding* (MMLU) tasks, several benchmarks have been developed. Examples include *Open LLM Leaderboard*[24],

[24] https://huggingface.co/open-llm-leaderboard

LMSys Chatbot Arena Leaderboard [25] or *HELM* [26] (Stanford). These benchmarks evaluate Generative AIs with adapted metrics on data corresponding to specific tasks such as answering multiple-choice questionnaires (MCQ), summarizing, performing sentiment analysis, retrieving information, reasoning, generating code. The metrics are aligned with the specific need as concordance (*exact match* for the MCQs or sentiment analysis), F1 score (for tasks for answering the questions without proposed answers), RED-N (summary), number of generated codes passing the unit tests (code generation), demographic representation (bias), proportion of toxic generations (toxicity).

4.5. Data Governance

To start with, it is worth mentioning that data governance in the construction of an AI model can be based on the framework of financial institutions, i.e., on the principles established by the Basel Committee with BCBS 239.

BCBS 239 defines standards to ensure data control, knowledge and, above all, quality[27]. Given the risks related to bias, the use of personal and sensitive data and the interpretability of data and the traceability of the data sources used, the requirements expected by BCBS 239 and the GDPR contribute strongly to the control of these risks.

AI systems based on incremental or online learning make data governance more complex. Examples include:

- Complexity in tracing and controlling re-training data due to the dynamic nature of learning. It is necessary to be able to implement controls on the fly to ensure quality (e.g., detection of aberrant, missing, invalid variables, etc.) before updating the parameters of the model.
- The impossibility of keeping the exact version of the model used to make the predictions. It may therefore be necessary to store explanations of the predictions (e.g., Shapley[28] values).

[25] https://chat.lmsys.org/?leaderboard
[26] https://crfm.stanford.edu/helm
[27] The quality criteria shall cover in particular the accuracy, integrity, completeness and timeliness of the data.
[28] Christoph Molnar, Giuseppe Casalicchio, Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. Joint European conference on machine learning and knowledge discovery in databases. Cham: Springer International Publishing, 2020. https://arxiv.org/pdf/2010.09337

4.5.1. Data collection

CHECKLIST OF DATA COLLECTION PRINCIPLES

A ***checklist* of data collection principles** should be established to identify best practices and questions around data collection and to ensure appropriate use of data.

This *checklist* must at least adopt the following recommendations:

- Consider the relevance and adequacy of the data sources used for the model,
- Ensure regulatory compliance on the use of the data considered (e.g., GDPR, *AI Act*, etc.),
- Specify the data warehouse and storage architecture to be used, their accessibility and the associated security,
- Ensure the freshness and availability of data by defining the frequency of collection and feeds,
- Define the controls to be implemented in order to ensure the quality of the data collected, in a sustainable manner over time,
- Document information about data collected in a data dictionary to make it easier to reuse.

Any additional element resulting from the principles of GDPR may be added to this checklist in order to ensure the compliance of the processing relating to sensitive and personal data.

ACCESS AND ACCREDITATION FORM TEMPLATE

In addition to the checklist, each data or data source must be authorized for access and use by the data owner. **This access and accreditation form shall define in particular:**

- Conditions of usability and duration of use,
- The *Service Level Agreements* (*SLAs*) to which the data owner commits,
- Conditions of access and accreditation.

4.5.2. Knowledge of data

A good knowledge of data is necessary in order to control the risk of bias as well as the correct interpretation of the data used for training the models.

In order to centralize, identify and share information about data and its uses, **a data dictionary** is needed. It is highly recommended that organizations have a data governance tool to document dictionaries, usages, sources, and controls of their critical data.

DICTIONARY AND/OR DATA CATALOG

The data dictionary shall be used, inter alia, to explain and centralize the data used by the AI system:

- Description of the data, their technical locations, and the format of the data structure to be used,
- The *Golden sources* on the priority data defined by the entity,
- The owners of each data in charge of ensuring its quality,
- The rules for the use of each data item and the associated regulatory constraints (GDPR, AMF storage period, etc.),
- Current uses of these data,
- Depending on the entity's maturity level, *Data Lineage,* that is, the traceability of the life cycle of the data explaining its production path (from the *golden source*) to its loading into the various decision warehouses.

4.5.3. Data quality

Data quality is an imperative in order to avoid bias and ensure the model's effectiveness.

However, as data moves through various databases and applications before it is used, it is essential to ensure its quality throughout its lifecycle with the implementation of regular controls: from the *golden source* in operational bases to the decision-making warehouses. The development of AI systems makes it possible to identify the data that will be critical during their production phase, to define the expected level of quality, and to put in place the necessary controls to monitor quality. This deployment of controls must be based on the organization's existing governance (in particular by capitalizing on the framework put in place for compliance with regulatory requirements such as BCBS 239).

DATA CONTROL METHODOLOGY

To this end, defining a **data control methodology** allows each entity to define:

- Data quality managers (Producer vs Consumer entity),
- The quality criteria to be checked: completeness, timeliness, uniqueness, compliance, integrity, accuracy, and consistency,
- Their origine and operational deployment: we will be able to follow a "*What? Where? When? How*?" to implement controls,
- The system for monitoring results and managing quality gaps.

The operational deployment of these controls should then be regularly monitored.

4.5.4. Data pre-processing

Good data preparation is fundamental to ensure the reliability, model efficiency and regulatory compliance of AI models. In particular, it helps to control data biases and paves the way for increasing/improving/correcting data in case of missing or erroneous data.

BEST PRACTICE GUIDE

To do this, it is recommended to rely on **a <u>best practice guide (or even a procedure in some regulatory cases)</u>** used to prepare and monitor data before they are used for modeling. The guide should contain the following steps:

- Description of the domain and dataset, using the data dictionary to:
  - Identify the objective of the project and the key issues to be resolved,
  - Describe data sources, their provenance, and their relevance,
- Data Collection and Integration:
  - Gather data from various sources, if necessary,
  - Ensure proper integration and merge datasets in a consistent manner,
  - Define data controls ensuring their quality at the time of integration into the tables (see BCBS 239 quality criteria such as completeness, consistency, accuracy, etc.),
- Data Cleanup:
  - Process missing values (deletion, imputation, etc.)
  - Correct errors and inconsistencies in the data (typos, outliers, etc.),
- Data Mining and Analysis:
  - Carry out an exploratory analysis (descriptive statistics, visualizations, etc.),
  - Identify potential trends, patterns, and anomalies,
- Data transformation and standardization:
  - Transform the data to make them usable (encoding of categorical variables, standardization, default value rules, etc.),
  - Resize data, if necessary (dimensionality reduction, etc.),
- Selection and creation of features:
  - Identify and select the most relevant features for the model (cf. 4.5.5 Data augmentation),
  - Eliminate redundant or uninformative features,
- Division of the Dataset to analyze and test the model:
  - Split the dataset into training, validation, and testing sets,
- Documentation of the dataset facilitating knowledge and auditability of the model:
  - Document the data preparation process, in particular data transformation policies,
  - Create and maintain metadata describing the dataset,
- Security and Confidentiality:

  - Ensure compliance with data privacy regulations (e.g., GDPR),
  - Anonymize sensitive data, if necessary,
- Data Backup and Versioning:
  - Regularly back up the dataset,
  - Use versioning systems to track changes.

In addition, all processes modifying the datasets should be traced in order to ensure a record of the modifications made.

It should be noted that most Data Science solutions include data preparation modules such as Alteryx, Dataiku or RapidMiner. Otherwise, some database languages also offer libraries for preparing data such as DBT or Pandas.

4.5.5. Data augmentation

FEATURE-STORE[29]

As mentioned above, "*features*" are essential to improve datasets and models: they "augment" the data by including data assessed as important by the experts (for example, the moving average over a period spent in *trading*); the use of synthetic data is another augmentation technique that we will not expand on in this document. Having a ***feature-store*** makes it possible to centralize, standardize and share these "*features*" in order to ensure their quality and reusability.

In addition, a *feature-store* optimizes model creation processes by providing quick and easy access to already prepared and validated features, reducing the time and effort required to prepare data.

FEATURE-STORE USAGE FRAMEWORK

In order to have a sustainable and easily exploitable *feature-store*, it is recommended to define **a framework stipulating the rules and best practices for its use**. Among these, it is important to specify in particular:

- The standard of the data and *features* to be stored (definition of data schemas, use of ETL pipelines (*Extract*, *Transform*, *Load*) to standardize the data, etc.),
- Access rights and authorized uses according to user profiles,
- Accessibility and interoperability with data science or development tools (API usage, authorized programming languages, etc.)

[29] See. §6. Glossary

- Metadata associated with *features* (*feature* versions, creation source, change history, *feature* users, etc.)

4.6. Modeling

4.6.1. Model construction

MODEL CONSTRUCTION STANDARD

A procedure or standard is a document that details the steps involved in building a model. It allows for a better homogeneity of the work (facilitating the recovery in case of departure of the data scientist or audit) and ensures that all key steps are covered by the data scientists.

DOCUMENTATION TEMPLATE

The Model Construction standard procedure can be supplemented by a *template* for model documentation, which makes it possible to standardize the information required to ensure that the relevant elements of the construction are explained and detailed. It also simplifies the task of reading and understanding model documentation for independent review teams (see *Model Risk Management*).

PIPELINE CONSTRUCTION TEMPLATE

Another important way to improve the traceability and transparency of modeling is to use a pipeline construction *template*, which allows for the decomposition of each step and flow of information, thus facilitating the understanding of the AI system.

BENCHMARKING TOOLS

Setting performance targets for the model is an important step: which performance metric should be chosen based on the use case, and what performance should the AI system achieve to be viable? Benchmarks can be:

- The existing process that the AI system seeks to replace.
- Simple rules.
- Human performance.
- AI systems based on different methodologies.

The use of Auto Machine Learning (AutoML) tools makes it possible to industrialize the creation of alternative models (called *challenger models*) to support modeling choices and select the best model for a given use case and constraints (e.g., performance metrics to be optimized, constraints to be satisfied). An AutoML tool typically ranks different models according to a performance metric chosen by the modeler as its primary outcome.

Available AutoML tools include: H2O, Flaml, Autosklearn, Datarobot, Mljar.

N.B.: It is possible to create your own *challenger models* without using an AutoML tool.

VERSIONING TOOLS

When constructing an AI system, it is recommended to keep the different versions of the models developed with their hyperparameters and associated performance. MLflow is an example of a tool for recording experiments and model configurations, comparing performance and parameters.

Other versioning tools (e.g., GitHub, GitLab, SVN) allow for collaborative code management and storage. The changes are thus recorded and traceable.

4.6.2. Model evaluation

PERFORMANCE MONITORING TABLE

The first step in evaluating a model is to establish a list of indicators to be optimized or satisfied:

- The **choice of a performance metric** to be optimized according to the use case (e.g., *accuracy*, F1 score, precision, reminder for classification, or R^2 for regression). For Generative AIs, the choice of the performance metric will depend on the general capacities used. Examples include the *exact match accuracy* for Q&A or *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE-N) for summary tasks. This last metric assesses the similarity between an automatically generated summary and a human reference summary by comparing the sequences of n consecutive words (*n-grams*) between the two texts.
- The **definition of constraints to be satisfied**. They can be technical (e.g. CPU or GPU utilization, latency, model size, number of *tokens* generated for LLM), ethical (e.g. *fairness*), related to transparency (e.g. ability to understand model decisions and ensure that the prediction is made for the right reasons), robustness or quantification of predictive uncertainty.

These indicators will make it possible to establish a ranking of the models, also called a *leader board*.

BIAS DETECTION TOOL

Some tools can detect and address biases (e.g., AIF360, Fairlearn), others are specific to the production of explanations (e.g., SHAP, AIX360, interpretML) or to the measurement of uncertainty (e.g., MAPIE).

Indicators can be analyzed globally or on specific regions of the *dataset* to identify potential problems on sub-populations. For example, it may be interesting to explain the model's erroneous predictions using XAI [30] techniques to highlight potential biases.

[30] Explainable AI

4.6.3. Model Selection

Candidate models assessed in the previous step that do not meet the constraints or are deemed to be underperforming are eliminated. Among the remaining models, stakeholders in the development of the AI system need to agree on the list of key selection criteria (e.g., performance, degree of interpretability, absence of bias) and potentially make trade-offs (e.g., performance and explainability) depending on the use case. The business experts' analysis of the relevance of the most important explanatory variables for predictions is key in the model selection process. Technical feasibility related to the implementation in production is another important criterion in the selection of the model. For example, some models require the use of GPUs or *cloud* solutions to function properly or with an acceptable latency. Technical limitations such as the *hardware* limitations of the production environment may reduce the choice of modeling techniques.

VALIDATION GOVERNANCE OF THE CHOICE OF AN AI MODEL

It is good practice to establish a validation governance for an AI system. It allows for the definition of the decision-making modalities and the designation of the persons responsible for the approval. The review by a team independent of the one in charge of modeling makes it possible to confirm or question the choice of the model and its construction.

In financial institutions, this governance to choose the model to be put into production could be based on the existing risk management framework for models. In addition, for more mature companies, it is recommended to integrate validation as much as possible into the model development process.

4.7. IT

4.7.1. Model Deployment

After development, a model is usually deployed in a production environment. This deployment first and foremost is an IT project and must follow the best practices and requirements of the company.

AI MODEL DEPLOYMENT STANDARD

In the context of artificial intelligence systems, MLOps (*Machine Learning Operations*) is a good example of standardized practices aimed at deploying *Machine Learning* models in production in a reliable and efficient manner, and generally covers all phases of development, production and model monitoring.

For example, MLOps covers:

- *Data Engineering*: collection and preparation of data and *features.*
- Model building: model modeling and training.
- Model validation: verification of model performance and other risk criteria.
- Model deployment: integration of the model into the IT environment and business processes.
- Monitoring: continuous or periodic verification of the acceptance criteria of the model such as the performance of the model.

For each phase, various degrees of automation are possible:

- Manual process: the steps are performed manually and the transition from one phase to another is also done manually.
- Automation pipeline: Automation steps are developed within "pipelines", e.g., variable selection pipeline, training pipeline, etc.
- CI/CD: the entire chain can be automated through "*Continuous Integration*" tools i.e., automation and planning of the development and testing phases of a model, and "*Continuous Deployment*" tools i.e., automation of controls and the deployment of a model.

Automation of this process is only possible if:

- The end-to-end process is clear and documented.
- Roles and responsibilities are clearly defined for each actor and at each phase.

- IT teams are involved from the earliest stages of the project to ensure a smooth transition to the production IT environment and anticipate potential integration and deployment challenges.
- The costs of the AI model deployment are clearly assessed, as scaling up such projects often generates unexpected costs, which becomes a bottleneck (underestimation of the resources needed, inefficiencies in resource management, lack of support, etc.) at the end of the project.

AI MODEL DEPLOYMENT TEMPLATE

In order to facilitate the deployment of an AI model, the following information should be formalized:

1. **Model description**: A detailed description of the AI model, including its purpose, the data and the algorithms used.
2. **Model performance indicators**: A list of performance indicators to assess the model accuracy and quality during testing and validation phases.
3. **Model deployment procedures**: Step-by-step instructions on how to deploy the model in a production environment, including the infrastructure used, software configurations and data integration pipelines.
4. **Model monitoring procedures**: Details on how to monitor the model in production, particularly the performance indicators, the tools to be used and the frequency of monitoring.
5. **Model maintenance and update plan**: A plan to update and maintain the model over time (continuously or periodically).

AI MODEL MANAGEMENT AND DEPLOYMENT TOOL

The orchestration of these stages is increasingly automated with tools that facilitate both the deployment and the monitoring of models that have gone into production. Several tools are available:

- MLflow assists *data scientists* in the development and deployment of models, as well as the life cycle management and evaluation of models, the monitoring of the components created (features, pipelines, configurations, etc.), up to the monitoring.

- Other tools exist such as Kubeflow, Comet, etc., as well as tools natively integrated within cloud service providers to manage the services and applications offered.

AI MODELS INVENTORY

An inventory of AI models is essential to understand where they are deployed and used, allowing a more transparent use of this technology for business lines and customers.

Having an inventory also makes it possible to monitor the evolution of the use of AI in a company and the maturity of the teams building these models. In this way, a clear and transparent vision will also help to meet requirements for risk management and monitoring.

4.7.2. Model Maintenance

MODEL MONITORING GUIDE

This guide supports the first Line Of Defense (LOD) teams in setting up the model monitoring framework. It lists the roles and responsibilities, for example who will be in charge of calculating the monitoring metrics or who will decide to retrain the model. The monitoring guide also lists the dimensions to be monitored (e.g., data distribution drift, data quality, performance drift, incidents, IT metrics). It thus proposes and defines examples of common metrics used to monitor these aspects. It suggests, in a generic way, examples of possible actions depending on the levels of these metrics. Finally, it lists all the other important elements to be defined according to the materiality of the model, such as the monitoring frequency or the required data infrastructure.

A Monitoring Protocol Template can therefore be useful to standardize the methods of the teams responsible for monitoring. This also makes it possible to speed up the monitoring implementation process.

MONITORING TOOLS

MLflow allows the tracking of the experiments, such as the evaluation of a model in production over a period of time with the display of the calculated performance metrics. MAPIE is a Python library that quantifies the uncertainty of a model's

predictions. Increasing uncertainty may be a sign of a distribution drift in the input data.

In addition, as when building a model, explainability tools can generate explanations on demand: e.g., SHAP, AIX360.

Finally, in order to monitor carbon emissions, Python packages such as CodeCarbon or CarbonTracker can be integrated into the inference pipeline.

4.8. Transfer to the business line

The transfer of an AI model to the business line is made during its operational implementation. This phase includes a set of elements that are not specific to AI, such as documentation, training, change management, deployment of controls and monitoring of uses by the business.

As a first step, key people impacted by the project must be identified. For an internal model, this will be the direct users of the model and their reporting lines.

Users must have clear instructions for integrating the model into the business process, to become autonomous in interpreting the model results and knowing the limits of its use. Modelers can also provide use cases and guidance for the continuous maintenance to the business.

The documents to be provided to ease the transfer of an artificial intelligence model to a business line may include:

DETAILED DOCUMENTATION

Overall, it is necessary to be able to provide detailed documentation of the model which includes information on the model's design, its deployment in the information system, the processes followed for technical and IT validation as well as information on security and confidentiality (including protection of sensitive data), and information on the limits of use of the model. This documentation may be divided into several documents if necessary.

This documentation must be understandable by non-data scientists: details on the model architecture, the type of algorithms chosen and the key parameters. In addition, documentation on the data used may also be provided to the business line, including key information on the datasets used for training, their source, quality, and pre-processing treatments performed.

The documentation shall include **performance metrics** (model evaluation results) measuring its accuracy.

MODEL MAINTENANCE AND USER GUIDE

This guide is intended to be shorter than the detailed documentation and is targeted at end-users. It contains instructions on how to integrate the model into existing business processes, on model usage limits, and information on model maintenance, including the metrics that are tracked and the associated monitoring process.

TRAINING KIT

The aim is to develop a training kit with concrete examples showing how to make the most of the model in real-life scenarios, and the main characteristics of the model (summary of the elements already mentioned).

SUPPORT AND FEEDBACK PROCESS

Finally, to support the operational use of the AI solution, a clear process of roles and responsibilities must be defined, which may include the following information:

Support Contacts: contact information for technical support in case of questions or problems.

A user feedback template that allows users to report potential weaknesses of the AI system in the template

4.9. Transversal phases

The transversal phases (Figures 4 and 5) aim to ensure the security of the AI application (model and data). Classic/common cybersecurity tasks protect AI systems in the same way as any other computer systems by securing their access or protecting their data. However, AI and cyber communities start realizing that AI can be attacked in different ways: indeed, many AI-specific attacks (adversarial attacks) are beginning to be identified[31], revealing new risks. For example, **data poisoning** attacks of the training dataset **will degrade the model's performance; evasion attacks** (or manipulation of Generative AI) will alter the data in production

[31] Apostol Vassilev, Alina Oprea, Alie Fordyce, Hyrum Anderson. Adversarial Machine Learning. NIST. January 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

in order to modify model predictions; **confidentiality attacks** (or more generally **exfiltration attacks**) will exploit the AI system in production to extract data (in particular personal data); **abuse violation attacks** (for Generative AI) will redirect the system by prompt injection to promote hatred or discrimination, disseminate malicious code, etc.

Mitigation options are proposed in the NIST document, and in the MITRE[32] document. However, in many cases, especially for Generative AI[33], there is not yet a complete or foolproof solution to protect the models.

The best method would of course remain to design and develop AI systems secured "by design", which is what the ANSSI recommends along with the major global cybersecurity organizations[34]: unfortunately, these recommendations are still not very operational compared to the risks, which are growing rapidly.

4.9.1. Securing models and data

IT SECURITY FRAMEWORK

Securing a model is done at all stages of the AI model life cycle as the sources of vulnerabilities or potential attack surfaces are multiple:

- First, a **comprehensive IT** (or cybersecurity) **security framework** needs to be put in place, which is generally the case for banking institutions already subject to various regulations. This includes ensuring the robustness of components (hardware and software) used during the AI models' lifecycle, for example, preventing unauthorized access to data used to train a model, to production platforms, or to network equipment.
- The development of AI models heavily relies on **open-source components**, which can themselves be attack vectors by carrying vulnerabilities or viruses. The source of these components should be verified, and scans should be performed to detect possible viruses or vulnerabilities in the libraries.
- Similar to open-source components, **open-source AI models** can be used to develop a new model or be integrated into software. The verification of the

[32] https://atlas.mitre.org/matrices/ATLAS
[33] https://owasp.org/www-project-top-10-for-large-language-model-applications/
[34] ANSSI Safety recommendations for a generative AI system. Guide. 29 April 2024. https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative
UK National Cyber Security Center. Guidelines for secure AI system development. 2023. https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

source, and the detection of viruses is necessary in order not to introduce disruptive elements into the information system.

- Additional checks can be performed at the time of the model usage to detect **"adversarial" attacks**, such as data poisoning, prompt injection or manipulation of input data to disrupt the values generated by an AI model. These types of attacks are increasing and are receiving particular attention from cybersecurity and research teams today.

Furthermore, checks on these various safety aspects should be carried out regularly to identify ineffective controls and thus strengthen the same control system.

4.9.2. Personal data protection

Training an AI model requires a varying amount of data. It is necessary to ensure that this data can actually be used, i.e. that the data is lawful and does not present an associated legal problem, for example, in some cases, obtaining the consent of the individuals in order to be able to use their personal data. In section 4.5 we have covered the risks related to data governance, we focus here on personal data.

Data protection policies differ across regions and countries. The following is a non-exhaustive list of existing policies:

- **General Data Protection Regulation** (GDPR) allows the direct or indirect collection of personal data, under conditions of collection, information, or consent of persons.
- The "***EU-U.S. Privacy Shield Framework***" is a set of principles of personal data protection to which companies established in the United States of America are free to adhere. Companies established in the European Economic Area may transfer the personal data they process to US companies appearing on the "*EU-U.S. Privacy Shield Framework*" list, in the same way as transfers to countries recognized as "adequate" by the European Commission operate.
- The "***Act on the Protection of Personal Information***" in Japan is considered compatible with the European GDPR.

More and more countries are implementing similar data protection regulations, but compared to Europe, only a small fraction is considered "adequate." For other countries, transfers of personal data must be governed by transfer tools or contracts.

FRAMEWORK FOR CONTROLLING THE USE OF PERSONAL DATA

The identification of the impacts of regulations on the AI model must be carried out upstream and during the development of the model, in particular with IT and Data Protection Officers, or any other function that can respond to these issues.

4.10. Risks specific to Generative AI

In this section, we briefly discuss the specific risks associated with designing solutions based on Generative AI (note that experience on these topics is much less developed than for widely deployed predictive AI systems). Overall, we identified three main risk categories.

First, **the risks associated with the data** in which we find:

- **Risk of intellectual property infringement**, on the training data, but also on the generated content (copyright). It should be noted that this risk is reinforced using pre-trained models, for which training data might be unknown.
- **Risk of reproduction and propagation of biases** present in the training data.
- **Data confidentiality issues**, possible data leakage, lack of freshness of training data (~model staleness).
- **Lack of transparency**: training data, explanation of generated text (RAG).

The second category concerns **the model** itself and the possible errors and deviant behavior. We identify:

- **Risk of hallucinations**: generation of plausible but false content.
- **Generation of toxic or harmful content**: need for guardrails.

Finally, we have grouped in a third category **the cross-cutting risks** that are largely facilitated by Generative AI as well as the possible negative effects on society:

- **Cybersecurity vulnerabilities** regularly identified (not yet mature).
- **Facilitating malicious cybersecurity activities**: malicious use, jailbreak attempts, prompt injection.
- **Increased regulatory risk**: generation of inappropriate texts (compared to specific regulations such as customer protection).
- **Environmental impact**, whether on energy or water consumption.
- **Influence on sovereignty**: dependence on third parties (*LLM provider*).
- **Economic and societal impacts**: medium-term effects on the world of work, on certain industries.

4.11. Summary of the operational tools

The deployment and use of the tools described above within an organization can be accelerated using an "off-the-shelf" software solution on the market. These solutions are of two types:

- **Solution as a platform**: A suite of tools available as a platform for the organization to use. The platform is supported by its editor and is accessible with a subscription or a contract. Most of the platforms will be offered as a SaaS service and accessible from a web browser. Some editors offer additional services (training, consulting, support) to enlarge their offer.
- **Solution as a tooled consulting service**: offered by service providers, these services are intended to support an organization in the production of its AI systems. In order to carry out their activities, providers can rely on tools, without necessarily giving customers access to them. These tools may be the property of the provider (developed by him, for his internal use), or acquired (subscription to a third-party service, or purchase of intellectual property).

The Hub France IA, together with the Banking and Auditability working group, sought to evaluate the existing offer in France: to this end, we developed a methodology for analyzing offers and interviewed a panel of solution editors. The work carried out is described by Hub France IA in the technical[35] note published in June 2024. This note presents the findings of the solutions analysis and is proposed as a guide in decision-making for choosing a solution in response of an organization's specific needs. It also proposes a methodology for evaluating solutions.

This study, carried out within the framework of the Banking and Auditability Group, naturally reflects the expertise of the financial sector, which is particularly regulated concerning the model management. However, the results of the analysis may provide conclusions applicable to other sectors.

Firstly, we describe in detail the methodology used by the experts to identify, assess, and analyze the solutions. The results of the analysis are presented in the next chapter.

Note: The analysis of the solutions cited in this study does not reflect their absolute quality. This analysis illustrates the level of coverage of each solution compared to

[35] Hub France IA. June 2024. https://www.hub-franceia.fr/telecharger-le-pdf-operationnaliser-la-gestion-des-risques/

the business process described in this document, i.e., its ability to operationalize the IA Act compliance process. Thus, the overall score presented for each solution must be put in perspective according to its context of use.

The production of the analysis was organized in three stages:

1. Construction of the analysis grid to be used to evaluate the solution.
2. Selection and analysis of solutions using the analysis grid.
3. Data processing and retrieval via the technical note.

The exercise started in October 2023 and ended when the note was published. Hub France IA is the sole contributor.

Step 1: Build the analysis grid

The solutions are analyzed using a grid built from the production process of an AI system. Each analysis item represents one step in the process. The analysis consists, for each step, of assessing whether the solution studied proposes services relating to the concerned step.

For each step, a **coverage level** is expected:

- **Covered**: The solution provides the services to fulfill the step.
- **Partially covered**: the solution offers services that can contribute to fulfill the step OR future evolutions are planned on the solution to fulfill of the step in the near future.
- **Not covered**: The solution does not provide services to fulfill the step.

Step 2: Analysis of solutions

Note that this study does not propose an exhaustive referencing of solutions but relies only on a panel.

The analysis of the solutions is carried out in three successive stages:

1. The **identification**: identification of the solutions analyzed and referenced in the document.
2. The **demonstration**: the editor introduces its solution.
3. The **analysis**: filling in the analysis grid.

The identification of the solutions chosen is carried out by the Hub France IA membership community. This identification and subsequent selection of each solution was motivated by the following reasons:

- A member of the Hub France IA is already a user of the solution.
- The editor of the solution is a member of the Hub France IA.

Step 3: Data processing

Following the analysis phase leading to the completed and coherent grids, the Hub France IA processed the data to produce the results presented in the technical note.

These results are:

- **Quantitative**: calculated from the coverage levels of the analysis grid.
- **Qualitative**: based on the comments justifying the coverage of each step of the process, they provide a textual description of each solution.

A "radar" makes it possible in particular to position each solution in relation to its coverage of the needs of each of the phases of the production process of the AI system.

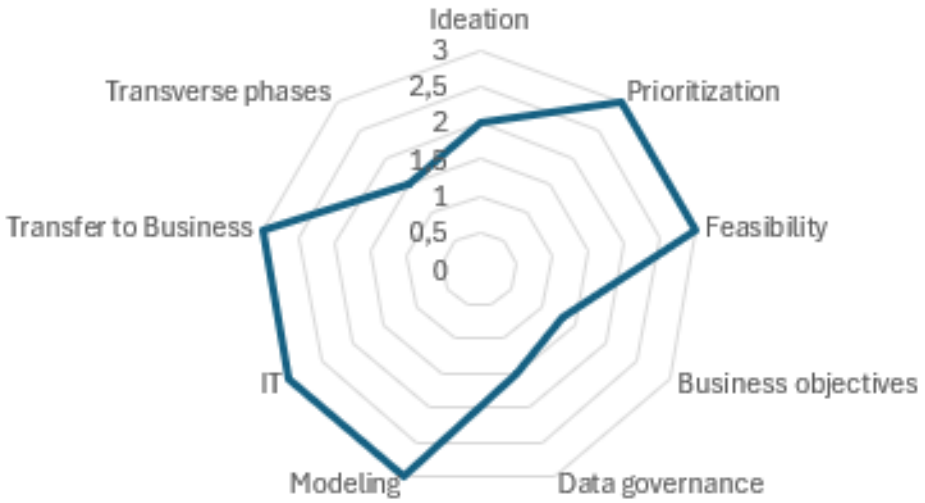

Figure 7 – Radar to cover a solution in the AI process phases

HUB FRANCE IA

5. Conclusion

5. Conclusion

The risk analysis we have just described reveals **many tools** to be put in place throughout the AI system implementation process. We have shown, certainly without being exhaustive, the importance of monitoring tools capable of supporting the process, which will be additional to the usual IT tools and will have to facilitate the compliance required by the AI Act.

The market is seeing the emergence of many **AI Act software solutions**, but, as we have seen through the solutions we have analyzed, the coverage of these solutions is still fragmented. No unique solution has yet emerged, and the market should of course be expected to strengthen in the coming years. Furthermore, we did not seek to align the adequacy of these tools with **the legal obligations of the AI Act**, which was not yet published at the time of our work (neither were the harmonized standards and delegated acts). It is still difficult to predict how the AI Act will be implemented and impact the practice of risk management as the field is rapidly evolving.

Finally, the deployment of **Generative AI** in companies has just begun. It is already clear that it raises new risks, but there are still many legal uncertainties (copyright for example). This area should therefore be monitored in the coming years (months?) in order to put in place the necessary tools for risk control.

We hope that this work, carried out by the Banking and Auditability working group at Hub France IA, will provide useful elements for all sectors that will be confronted with the AI Act.

6. Glossary

AutoML	Also called automated *machine learning*. Automates tasks for developing a *Machine Learning* model, such as preparing data, selecting variables, training, and so on
BCBS 239	*Basel Committee on Banking Supervision's standard number 239*: Principles for risk data aggregation and risk reporting
ChatGPT	Chatbot developed by OpenAI, based on a Large Language Model
Chunk	Block of information extracted from a larger dataset
CI/CD	*Continuous Integration/Continuous Delivery* or Continuous Integration/Continuous Deployment is a practice of automatically integrating code changes into a repository, integrating, and testing those changes, and then automatically deploying them into production environments.
CNIL	*Commission Nationale de l'Informatique et des Libertés* French authority in charge of GDPR application.
CPU	*Central Processing Unit*: the main microprocessor of a computer
Dataset	Set of data used in one of the modeling phases
Deep Learning	A subset of *Machine Learning* based on the use of so-called deep neural networks, i.e., using multiple layers of neurons
Embedding	An *embedding* is a large vector representation for an object such as a word, document, or image. These representations contribute to the calculation of attention in the Transformers models and make it possible to know which words are semantically close
ETL	*Extract, Transform, Load*: software to collect data from multiple sources
Feature	Explanatory variable at model input

Feature store	It is a database that allows the efficient storage, sharing and management of "*features*", to facilitate their reuse between projects and models, for training or in production
Few-shot learning	The integration in the *prompt* of Generative AI of some examples of the task to be performed is called *few-shot learning*
Fine-tuning	The *fine-tuning* of a pre-trained Generative AI consists of having it perform additional training on labeled data specific to a particular task or field to improve its performance
GPT	*Generative Pretrained Transformer*: a family of *Large Language Models* developed by OpenAI
GPU	*Graphics Processing Unit*: processor specializing in image rendering, 2D/3D image processing, and mathematical calculations
Guardrails	These are protections that control the inputs and outputs of a Generative AI to reduce the risks related to its use
Hallucination	False, inaccurate, or inconsistent information created by a Generative AI
Human-in-the-loop	Human intervention in decision-making
Generative AI	A subset of *Deep Learning*, designed to produce content, whether text, image, audio, or video, from input data (called *prompt*), itself text, image, audio, or video.
Pre-trained Generative AI	Generative AI is "pre-trained" if it has been trained on very large volumes of data, allowing it to acquire general knowledge and the bases to generate relevant content
KPI	Key Performance indicator
Large Language Model (LLM)	A type of Generative AI capable of generating and analyzing text (e.g., natural language, programming language, etc.)
LOD	*Line Of Defense*: Level of control making up the internal control of an institution

Machine Learning	Machine learning from a dataset
MLOps	A set of practices that aims to deploy and maintain *Machine Learning* models in production reliably and efficiently
Model Owner	A key player responsible for ensuring that the development of the AI model, its implementation, its use, and its monitoring over time comply with the bank's policies and procedures
Override	Human decision to override and change a result given by a system
POC	*Proof Of Concept*. Refers to an achievement intended to demonstrate the feasibility of a project
Prompt	The *prompt* is the natural language instruction or request provided to Generative AI for the purpose of obtaining a response
RAG	Retrieval *Augmented Generation*: increased generation by retrieving information from a knowledge base that was not used during Generative AI training
GDPR	General Data Protection Regulation
ROUGE-N	N-grams based callback between a generated summary and a set of reference summaries
SLA	*Service Level Agreement*: A service contract between an IT provider and a customer
IT Security	Information Technology Security, see International Standard ISO/IEC 27001 and National Information Systems Security Authority (ANSSI)
Supervised Learning	A *machine learning* method where a model is trained on a set of labeled data or with target values, i.e., each entry is associated with a label or target value, thus enabling the model to learn to predict labels or target values on new data.
Temperature	Temperature in Generative AI is a parameter in the model used to manage the randomness of generated text. Temperature value close to zero will generate almost identical text for each generation, while a high temperature value close

	to or above 1 will generate text with more creativity or variability.
Token	A subset of a word that is a *Large Language Model* processing unit
Unsupervised Learning	A *machine learning* method where a model is trained on a data set without labels or target values. The model attempts to discover relationships between data, for example to group (*clustering*) or reduce the size of the problem
Zero-shot Learning	The *zero-shot learning* consists of feeding Generative AI with a "dry" *prompt*, without example of the task to be carried out

7. Acknowledgements

Contributors

- **Audrey Agesilas**, Supervisor - Model risk Audit, Société Générale
- **Benjamin Bosch**, Manager - Model Risk Management, Société Générale
- **Thomas Bonnier**, Model Risk Manager, Société Générale
- **Pierre Dehaene**, Data & IA Strategist, La Banque Postale
- **Lea Deleris**, AIR Tech, RISK x Compliance, BNP Paribas
- **Jérôme Lebecq**, Data Science Coordinator, BNP Paribas
- **Cyril Nicolotto,** Project Manager, Hub France IA
- **Chloé Pledel,** Head of European and Regulatory Affairs, Hub France IA

Reviewers

- **Caroline Chopinaud**, Chief Executive Officer, Hub France IA
- **Françoise Soulié-Fogelman**, Scientific Advisor, Hub France IA

The last touch

- **Mélanie Arnould**, Head of Operations, Hub France IA

BNP PARIBAS

SOCIETE GENERALE

LA BANQUE POSTALE

A toolbox for managing risks of Artificial Intelligence systems

Banking & Auditability Working Group

October 2024

HUB FRANCE IA