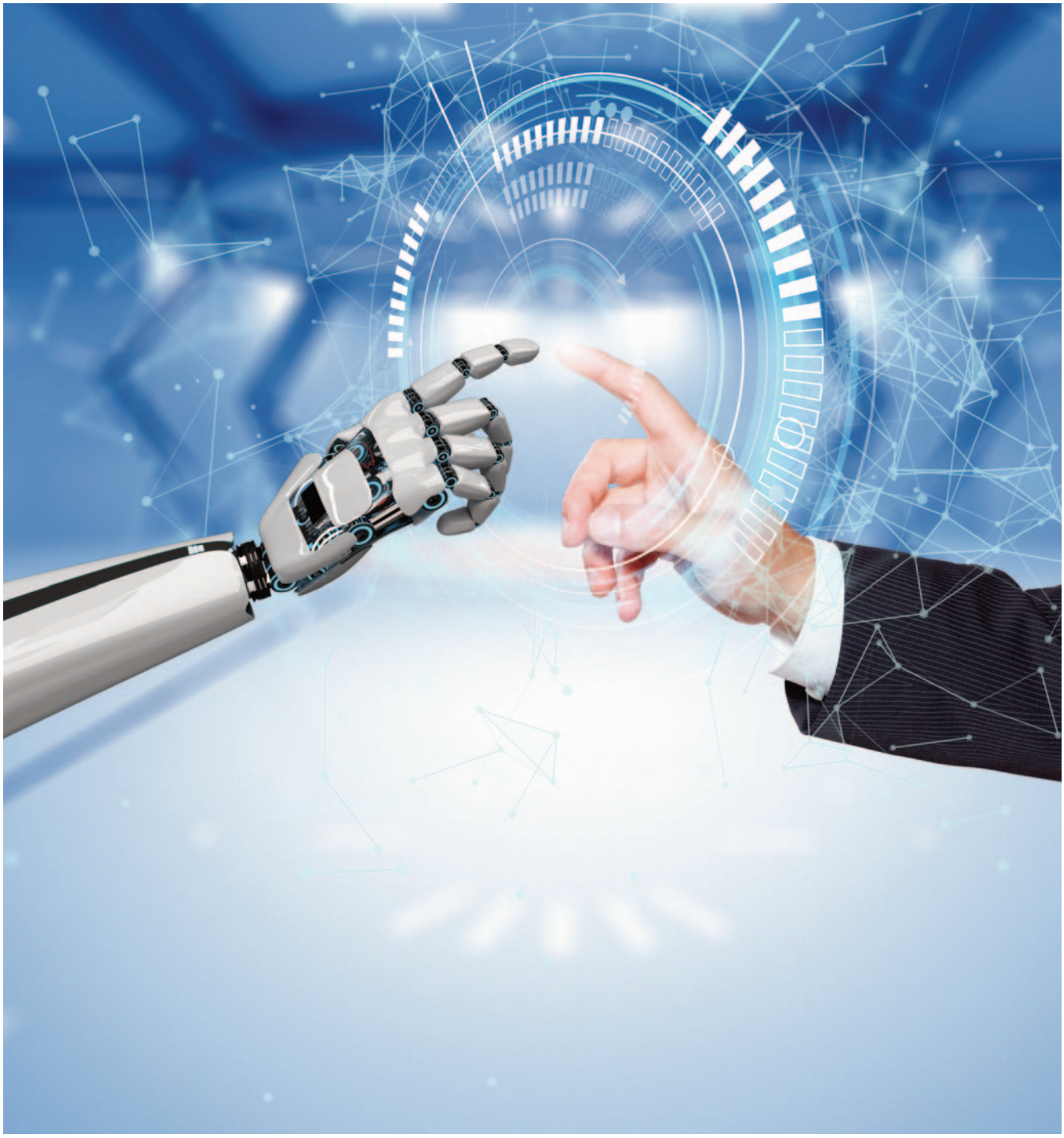
Context and rationale for XAI

"Understanding artificial intelligence is a challenge that requires enormous intellectual capacity; fortunately, we have artificial intelligence to deal with it".

GPT-4[16]



Context

One of the most notable features of digital transformation is that it is making a massive amount of structured and unstructured data from multiple applications available to all industries, for example:

- Retail data from purchase actions, transactions and customer feedback.
- Financial data from banking, investment and commercial sources.
- Social media data, including sentiment analysis and predictive analytics.
- IoT (Internet of Things) digital sensors that measure temperature, pressure and other environmental data.
- Health data, such as medical records, diagnoses, images and genomic information.
- Wearables, such as activity trackers, health sensors and smart watches.
- Speech recognition systems that allow machines to understand and respond to natural language.
- Satellites and other space-based sensors that provide weather and climate information.
- Intelligent surveillance systems using facial recognition and object detection.
- Autonomous vehicle sensors such as cameras, lidar, radar and ultrasonic sensors.

The availability of this data, coupled with the presence of enormous storage and computational processing capabilities at reduced cost, has driven an increased appetite for advanced modeling, manifested in the use of a wide range of machine learning techniques and the development of artificial intelligence (AI) in virtually all sectors and domains[17].

Although there is consensus that AI models generally provide greater predictive power than traditional models[18], they also introduce greater complexity and it can be difficult to interpret them and explain their results.

This generates risks associated with the use of these models, such as not properly understanding the model, the presence of inadvertent bias or the difficulty in determining whether the model is overfitted (globally or locally), which can result in insufficient generalization and potential errors in the decisions based on it, and as a consequence, lead to a lack of confidence in the model.

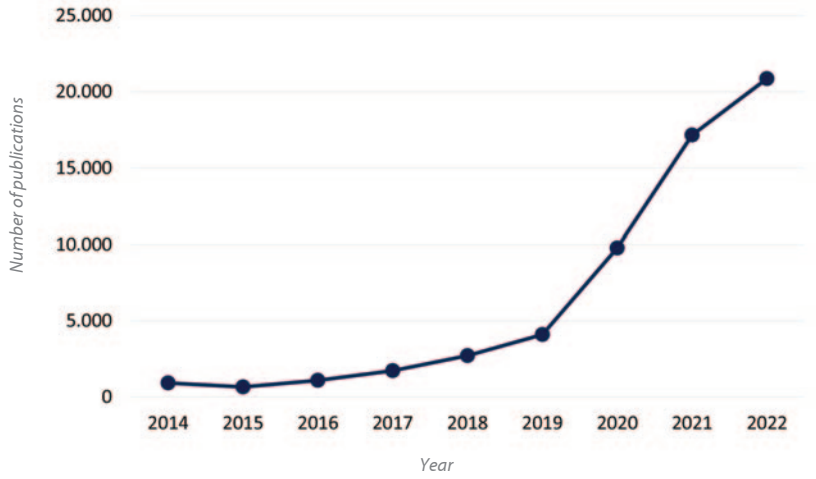
All of this brings up the question of whether it is possible to understand well enough the results that AI algorithms yield, especially when they impact critical decisions, such as medical diagnosis, autonomous driving or fraud detection, among many others.

[16] GPT-4, Generative Pre-Trained Transformer, a deep neural network designed by the OpenAI Foundation to perform natural language processing (NLP) tasks. In this case, GPT-4 was asked to "Come up with 10 clever quotes about artificial intelligence and how difficult and necessary it is to be able to interpret and explain AI models." The quote provided was the third one.

[17] Although there are differences, given the lack of consensus on their definition, the terms "machine learning", "machine learning (ML)", "artificial intelligence (AI)" and "advanced modeling" will be used interchangeably in this document. Likewise, the abbreviation "AI" will be used for "artificial intelligence", and the acronym "XAI" for Explainable Artificial Intelligence.

[18] LeCun, Y. et al (2015). Researcher at Facebook AI Research and New York University.

Figure 2. Number of scientific publications per year on Explainable Artificial Intelligence (XAI).



Definition

The XAI discipline is relatively new, and therefore there is not yet a settled doctrine that standardizes its terminology. Despite some notable efforts to define terms[19], the approach to XAI is either diverse (depending on the academic source consulted) or intuitive (more frequently in industry).

In any case, for most uses in practice it may be sufficient to define XAI as follows[20]:

> Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms. Explainable AI is used to describe an AI model, its expected impact and potential biases. It helps characterize model accuracy, fairness, transparency and outcomes in AI-powered decision making. Explainable AI is crucial for an organization in building trust and confidence when putting AI models into production. AI explainability also helps an organization adopt a responsible approach to AI development.

Relevance of XAI

One aspect on which there is consensus among academics and industry professionals is the growing relevance of XAI as a complementary discipline to AI.

Scientific publication analysis tools identify more than 77,000 articles on XAI between 2014 and 2022, and this trend is exponentially increasing, with more than 20,000 articles in 2022 alone (Fig. 2)[21].

Beyond academic interest, the attention XAI receives is explained by its ability to provide solutions to industry concerns around the use of AI (Fig. 3), including:

- **Regulatory requirements:** the obligation to comply with emerging regulations on the use of AI.
- **Lack of confidence:** the need to build confidence in the AI model and the results it delivers among users, validators and auditors, and ultimately the general public.
- **Potential misuse:** the desirability of avoiding misuse of the models due to lack of understanding of how they work, which can lead to costs and even penalties.
- **Reputational impact:** the prevention of reputational impacts for organizations due to model bias, discriminatory decisions, erroneous predictions by the model or inappropriate use.
- **Social or human impacts:** the prevention of harmful social or human impacts in critical uses such as AI for the diagnosis of medical diseases, judicial sentences, biometric identification, polygraphs, etc.
- **Other:** mitigation of other risks arising from lack of understanding about the model, such as cybersecurity, data protection, fraud, model risk, etc.

Despite all of the above, there are cases in which AI models do not need to be particularly interpretable, because their uses are not regulated, because they have no relevant potential impacts, or simply because they do not need to be interpreted, such as automatic movie and music recommendation systems, or algorithms that play chess, for example.

[19] Marcinkevics et al. (2020). Department of Computer Science, ETH Zurich.
[20] IBM (2022).
[21] Dimensions (2022).

Figure 3. Industry concerns that XAI contributes to solve

Regulation

XAI, therefore, is positioning itself as a discipline of growing relevance; and this is leading regulators and supervisors in different jurisdictions to establish regulations and guidelines for the appropriate use of AI, including model interpretability aspects.

In this context, possibly the most relevant regulatory references at the time of writing of this document are the following:

1. GDPR (European Parliament)

In Europe, the General Data Protection Regulation, which came into force in 2018, establishes citizens' "right to an explanation", according to which[22]:

> A data subject should have the **right not to be subject to a decision**, which may include a measure evaluating personal aspects relating to him/her, **which is based solely on automated processing** and which produces legal effects on him/her or similarly significantly affects him/her, such as the automatic refusal of an online credit application or online recruitment services where no human intervention is involved. [...]
>
> In any case, such processing should be subject to appropriate safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to **receive an explanation of the decision taken after such assessment** and to challenge the decision.

This has critical implications for the use of AI and may lead to questions about its feasibility. However, in the words of the European Parliament[23]:

> There is indeed a tension between the traditional data protection principles – purpose limitation, data minimization, the special treatment of 'sensitive data', the limitation on automated decisions– and the full deployment of the power of AI and big data. The latter entails the collection of vast quantities of data concerning individuals and their social relations and processing such data for purposes that were not fully determined at the time of collection. However, **there are ways to interpret, apply, and develop the data protection principles that are consistent with the beneficial uses of AI and big data.**

And this is in line with the fourth principle for the ethical use of AI established by the European Commission's High Level Group on Artificial Intelligence[24]:

> Explainability: processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected.

In any case, GDPR has a significant impact on the use of AI, in the sense that companies are legally obliged to be able to explain why an AI model has yielded a certain result, and this has critical implications on the design and interpretability analysis of AI models[25].

2. Artificial Intelligence Act (European Parliament)

The draft Artificial Intelligence Regulation or Artificial Intelligence Act (AI Act), published in 2021, is a proposal for the use of artificial intelligence in the European Union that aims to ensure a high level of trust in AI and its applications, while laying the groundwork for innovation. The Regulation establishes a regulatory framework for AI systems in the EU, and includes requirements for ethical development, transparency, security and accuracy. It also establishes a governance and oversight system for AI systems, as well as data protection and data governance rules.

[22] GDPR (2018), Cons. 71.
[23] European Parliamentary Research Service (2020).
[24] Ibid.
[25] In some European countries, the level of compliance of this type of AI (in particular, the so-called Large Language Models) with data protection regulations is being analyzed, and in some cases the use of some of these models has been provisionally banned.

As it is a Regulation, when approved, it will be directly applicable in the Union's 27 countries[26] without the need to be transposed into each country's legal system.

One of its key features is that it sorts AI applications into risk levels[27]:

- **Prohibited practices** is the highest risk category and systems falling under this category are totally forbidden. They include:
  - Real-time remote biometric systems that can be used for any type of surveillance, although exceptions apply for crime prevention and criminal investigations in law enforcement and homeland security contexts.
  - Social scoring algorithms that can be used to evaluate individuals based on predicted personal or personality characteristics leading to detrimental or unfavourable treatment of an individual.
  - Subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm.
- **High-risk AI systems** is listed in Annex III and is likely to constitute the majority of AI systems. These include:
  - Biometric identification and categorization of natural persons [...].
  - Management and operation of critical infrastructure [...] [e.g. traffic].
  - Education and vocational training [...].
  - Employment, workers management and access to self-employment [...].
  - Access to and enjoyment of essential private services and public services and
  - benefits [...], including creditworthiness assessment, credit rating or prioritization of access to such services (Note: this aspect applies to AI systems used in the financial services sector in particular).
  - Law enforcement [...].
  - Migration, asylum and border control management [...].
  - Administration of justice and democratic processes [...].
- **Low-risk (or limited-risk) IA systems**, covering systems that do not use personal data or make predictions that could affect individuals directly or indirectly, such as industrial predictive maintenance applications.

Regarding the interpretability of AI models classified as high risk, the AI Act establishes[28] in its Articles 13 and 14:

Art. 13. Transparency and provision of information to users

1. High-risk AI systems shall be designed and developed in such a way to ensure that their operation is **sufficiently transparent to enable users to interpret the system's output and use it appropriately. [...]**
2. High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users. [...]

Art. 14. Human oversight

1. High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use. [...]

[…]

4. The measures referred to […] shall enable the individuals to whom human oversight is assigned to do the following, as appropriate to the circumstances:
   a. **fully understand the capacities and limitations of the high-risk AI system** and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;
   b. remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias') [...];
   c. be able to correctly interpret the high-risk AI system's output [...];
   d. be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;
   e. be able to intervene on the operation of the high-risk AI system or interrupt the system [...].

As can be seen, the AI Act imposes restrictive conditions on the interpretability of high-risk AI models (Fig. 4), which will soon become mandatory throughout the Union. This is expected to trigger a significant number of initiatives to adapt to the Regulation, including more exhaustive documentation of models and their uses, the implementation of interpretability techniques, the development of model monitoring and alert dashboards, and a review of the full model development, validation, implementation and use procedure.

[26]Expected to come into force 20 days after its publication in the Official Journal of the European Union, and to be fully applicable 24 months after its entry into force.su entrada en vigor.
[27]Floridi et al. (2022).
[28]European Commission (2021).

3. Ethical Guidelines for Trustworthy Artificial Intelligence (European Commission)

In April 2019, the European Commission's High Level Expert Group on AI presented the Ethical guidelines for trustworthy AI[29], following a consultation process with more than 500 industry responses.

The Guidelines propose seven key requirements that AI systems must meet to be considered trustworthy, which in summary are: (i) human agency and oversight, (ii) technical robustness and safety, (iii) privacy and data governance, (iv) transparency, (v) diversity, non-discrimination and fairness, (vi) social and environmental well-being, and (vii) accountability.

Specifically with regard to AI model interpretability, the Guidelines establish the following as part of their transparency requirement:

> 53. Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and **decisions – to the extent possible – explainable to those directly and indirectly affected.** Without such information, a decision cannot be duly contested.
>
> **An explanation as to why a model has generated a particular output or decision** (and what combination of input factors contributed to that) **is not always possible.** These cases are referred to as 'black box' algorithms and require special attention.
>
> In those circumstances, **other explicability measures** (e.g. traceability, auditability and transparent communication on system capabilities) **may be required**, provided that the system as a whole respects fundamental rights.
>
> **The degree to which explicability is needed is highly dependent on the context and the severity of the consequences** if that output is erroneous or otherwise inaccurate.

As can be seen, the Guidelines point in the same direction: the requirement (which rises to the level of ethical necessity) that AI models be explainable.

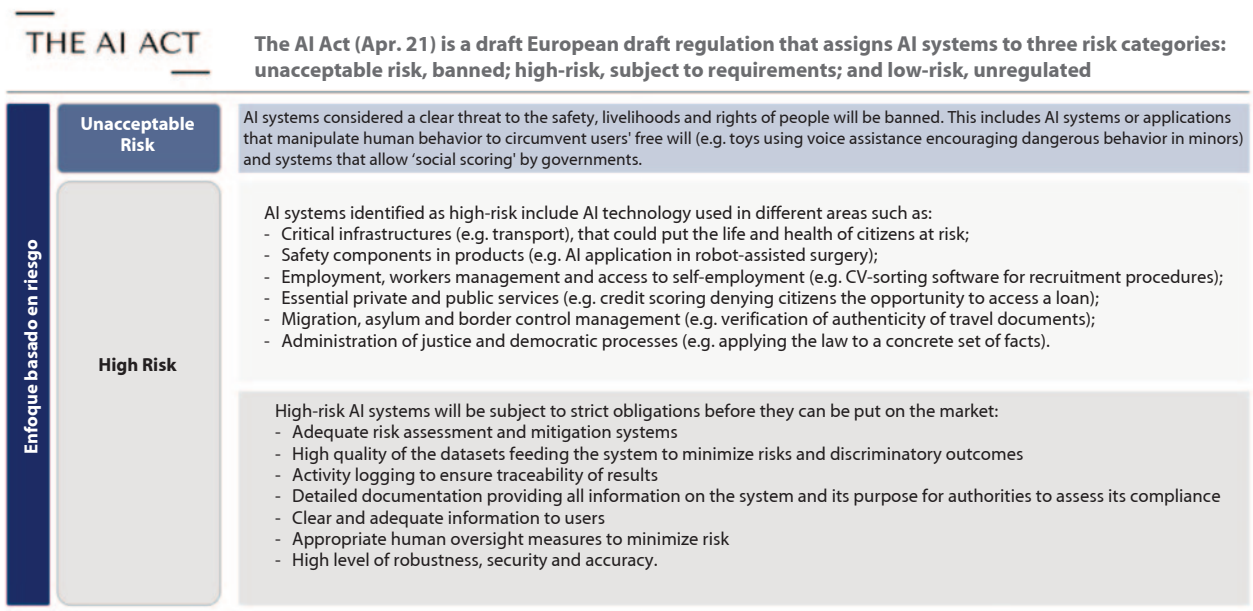
Likewise, what at first sight might appear to be a more relaxed requirement for AI model interpretability, since the Guidelines recognize that some AI models are more difficult to explain, in fact introduces an additional complexity: the need to classify AI models according to their interpretability risk and potential, in order to apply a greater or lesser degree of effort in their explanation.

Finally, the Guidelines are aimed at assessing the extent to which an AI model meets these seven requirements, and to this end propose a list of assessment criteria, which should be adapted to each specific case. With regard to explainability, the Guidelines formulate the following assessment criteria[30], which should be integrated with other assessment tools already available to organizations:

- Did you assess to what extent the decisions and hence the outcome made by the AI system can be understood?
- Did you assess to what degree the system's decision influences the organisation's decision-making processes?
- Did you assess why this particular system was deployed in this specific area?

[29]European Commission (2019).
[30]Ibid.

Figure 4. Application areas and requirements of the Artificial Intelligence Act.



- Did you assess what the system's business model is (for example, how does it create value for the organization)?
- Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?
- Did you design the AI system with interpretability in mind from the start?
- Did you research and try to use the simplest and most interpretable model possible for the application in question?
- Did you assess whether you can analyse your training and testing data? Can you change and update this over time?
- Did you assess whether you can examine interpretability after the model's training and development, or whether you have access to the internal workflow of the model?

4. Blueprint for an AI Bill of Rights (White House)

In October 2022, the White House proposed a Draft Artificial Intelligence Bill of Rights[31], driven by President Joe Biden and developed by the White House Office of Science and Technology Policy (OSTP), and accompanied by a handbook (From Principles to Practice) on how to implement it in practice.

The AI Bill of Rights sets out five principles or citizens' rights as they relate to AI, which are summarized as[32]:

- Safe and effective systems.
- Algorithmic discrimination protection.
- Data privacy.
- Notice and explanation.
- Human alternatives, consideration, and fallback.

Its fourth principle, on the explainability of AI models, includes that[33]:

> Designers, developers, and deployers of automated systems should provide generally accessible plain language documentation including clear descriptions of the overall system functioning, the role automation plays, notice that such systems are in use, the individual or organization responsible for the system, and explanations of outcomes that are clear, timely, and accessible.
>
> Automated systems should provide explanations that are technically valid, meaningful and useful to you and to any operators or others who need to understand the system, and calibrated to the level of risk based on the context. [...]

5. Principles on Artificial Intelligence (OECD)

The OECD Principles on Artificial Intelligence promote the use of AI that is trustworthy and respects human rights and democratic values. They were adopted in May 2019 by the 38 OECD member countries. They were the first such principles subscribed to by governments and include specific recommendations for public policy and strategy on AI.

Among other things, these principles state that "AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art […] to enable those affected by an AI system to understand the outcome"[34]. The OECD AI Policy Observatory, launched in February 2020, aims to help decision-makers implement these Principles.

6. Discussion Paper on Machine Learning for IRB Models (EBA)

Because of its relevance to the banking sector, the European Banking Authority's Discussion Paper on Machine Learning for IRB Models should be highlighted, published in November 2021 (Fig. 5).

The EBA's paper aims to analyze the relevance of possible obstacles to the implementation of machine learning techniques within the scope of the IRB approach to capital calculation in financial institutions, includes the challenges and potential benefits of using these techniques, and establishes certain principles and recommendations[35]. A central focus of the document is, logically, how to make the use of these techniques compatible with compliance with the European capital regulation (CRR[36]).

Regarding the interpretability of models, the paper addresses this under the "Concerns about the use of machine learning" section, and states[37]:

> The main concerns stemming from the analysis of the CRR requirements relate to the complexity and reliability of the ML models where the main pivotal challenges seem to be the interpretability of the results, the governance with a special reference to increased needs of training for staff and the difficulty in evaluating the generalisation capacity of a model (i.e. avoiding overfitting).
>
> To understand the underlying relations between the variables exploited by the model, several interpretability techniques have been developed by practitioners, [...][and] the choice of which of

[31] White House OSTP (2022).
[32] Ibid.
[33] Ibid.
[34] OECD (2019).
[35] See a detailed analysis in Management Solutions (2021).
[36] CRR: Capital Requirements Regulation, central regulation on capital in financial institutions in Europe.

these techniques to use can pose a challenge by itself, while these techniques often only allow a limited understanding of the logic of the model.

Beyond this, the document introduces the need to find a balance between model complexity and interpretability, and, unlike other regulations, it goes down to a more technical level when recommending the following to financial institutions:

a. Analyse in a statistical manner: i) the relationship of each single risk driver with the output variable, ceteris paribus; ii) the overall weight of each risk driver in determining the output variable, in order to detect which risk drivers influence model prediction the most. These analyses are particularly relevant where a close and punctual representation of the relationship between model output and input variables is not determinable due to the complexity of the model.

b. Assess the economic relationship of each risk driver with the output variable to ensure that the model estimates are plausible and intuitive.

c. Provide a summary document in which the model is explained in an easy manner based on the outcomes of the analyses described in point a. The document should at least describe:

   i. The key drivers of the model.
   ii. The main relationships between the risk drivers and the model predictions.

   The addressees of the document are all the relevant stakeholders, including the staff which uses the model for internal purposes.

d. Ensure that potential biases in the model (e.g. overfitting to the training sample) are detected.

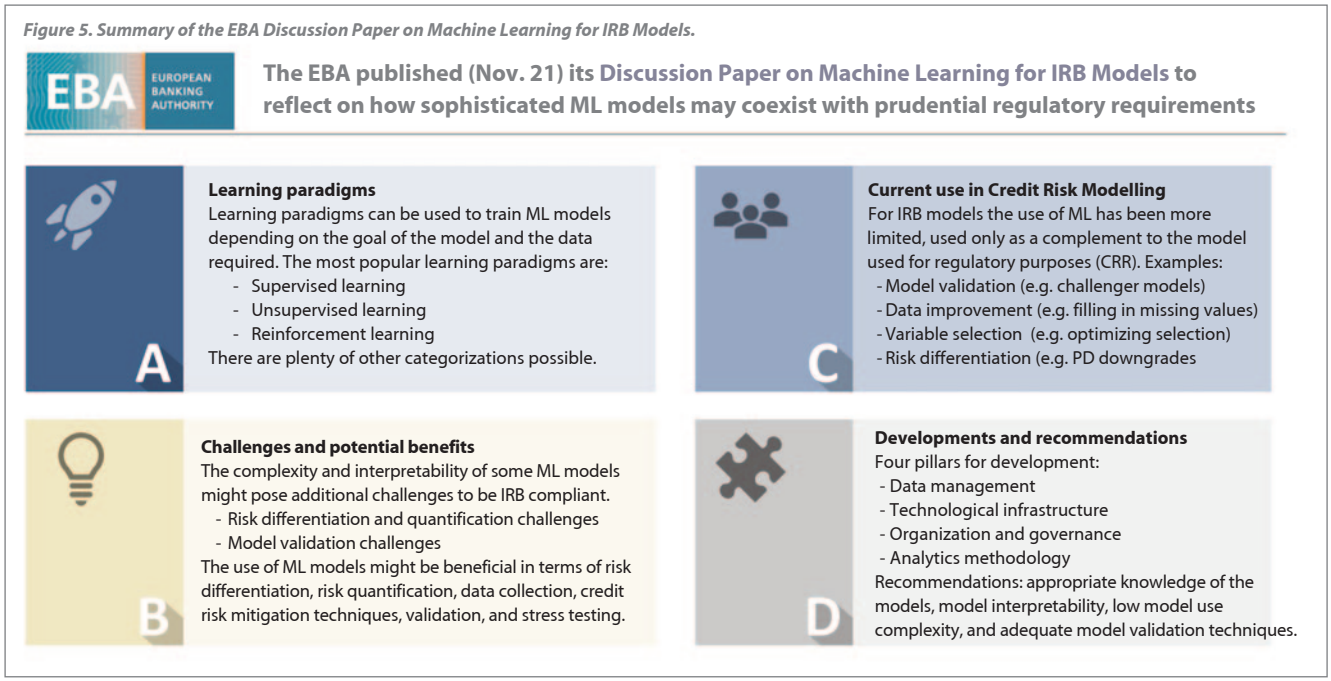
In practice, while the banking industry awaits the final version of the EBA consultation paper, most institutions using machine learning in their IRB models are already adapting their model development, monitoring and validation frameworks to ensure future compliance.

A common element in all regulatory references mentioned, as is apparent, is the need to provide an explanation to citizens on the use of AI, and to do so on two levels: the interpretability and transparency of the AI model as a whole, and the ability to explain specific model decisions, if required.

Beyond the regulatory references provided above, there are other publications, principles, guidelines and draft regulations in multiple jurisdictions that address AI model interpretability, both general and sectoral, and both regional and local to each country; the selection provided in this section includes those considered to have the widest scope and potential influence.

[37] EBA (2021).

Figure 5. Summary of the EBA Discussion Paper on Machine Learning for IRB Models.



Impacts on the organization and its processes

An essential principle of XAI as a discipline is that, beyond the development of specific explainability techniques or the construction of inherently interpretable models, this explainability and interpretability must be integrated into an organization and its processes.

Put into practice, this principle implies the development and implementation of an XAI framework, which can be structured into four elements:

1. Interpretability techniques for AI models
2. Integration into model risk management (MRM) processes
3. Technological support
4. Human factor

1. Techniques for AI model interpretability

The core elements of an XAI framework are the interpretability and explainability techniques, which can be summarized as having three aspects:

- Model design interpretability: this includes analyzing how the model would behave in different scenarios (e.g. adversarial attacks, extreme scenarios...), understanding how sub-models and model ensembles work, and integrating interpretability into the model design by applying constraints during model development.

- Interpretability of model results: this refers to detecting which variables influence the model prediction and how using both local (LIME, SHAP, etc.) and global interpretability (PDP, variable importance, surrogate models, sensitivity analysis); to assessing the economic sense of each variable (e.g. use case analysis of a representative data sample); and to ensuring that the model documentation correctly describes the model, including the input variables and their relationship to the results.

- Other aspects: ensuring detection of potential biases in the model (e.g. overfitting, biased input data, data errors) and periodically monitoring the model, especially when its scope changes or when it is applied to data other than development data.

Because of their importance, the main interpretability and explainability techniques will be discussed in the following section.

2. Integration in model risk management (MRM) processes

AI model interpretability is a feature that transcends development and impacts the entire model lifecycle chain, and thus the entire model risk management framework. A non-exhaustive summary of action required to incorporate XAI into a company's MRM framework would be as follows:

- **Governance:** update the organizational and governance framework to incorporate XAI; assess the impact of regulation applicable to AI models; update the model tiering system to address lack of interpretability as a major risk; update model inventory and inventory procedures to incorporate XAI elements (e.g. specific attributes for AI models).

- **Development:** update model development policies and procedures, as well as documentation requirements; evaluate fairness and bias, interpretability of inputs, design and results, data, supplier risk, predictive power metrics, limits to the use of AI models, etc.; perform sensitivity analysis of AI models to identify vulnerabilities; include specific tests for XAI in the development framework.

- **Monitoring:** update the model monitoring framework and complete it with specific XAI tests; review the thresholds and actions derived from non-compliance; develop early warning systems to detect changes in AI models; review compliance with model risk appetite; assess the need to develop an ad hoc monitoring module for dynamic learning models (i.e. that recalibrate automatically without human intervention).

- **Validation:** update the internal validation framework to detect potential risks associated with AI models and incorporate XAI tests; establish a cross-validation framework to ensure the quality of AI models; assess the impact of changes in the production environment on AI models.

- **Implementation:** update the model implementation process to incorporate tests specific to XAI features; update, if necessary, the technological platform to enable the implementation of AI models in production.

- **Use:** update procedures for the use of AI models to determine their suitability for the context in which they are to be used; review and complete user training on AI models; update protocols to detect potential situations of misuse or overuse of the models.

- **Audit:** implement an AI model audit framework to ensure proper implementation and use of AI models; establish XAI tests for auditing AI models; assess the adequacy of internal control systems to ensure the quality of AI models; analyze audit trails to detect potential risks associated with AI models.

Thus, the use of AI models entails a complete review of policies and procedures throughout the model's life cycle to incorporate the key components of XAI at the very least.

3. Technological support

The implementation of an XAI framework tends to start with departmental tools, and as soon as it reaches a minimum level of maturity, it requires professional technology solutions to support the interpretability aspects of AI models.

These solutions can be classified into two groups:

- Interpretability: development of systems that implement interpretability techniques in a standardized and homogeneous way. They should allow model interpretation to be performed in a manner that is automatic, easily configurable and ensures high quality, incorporating the most common techniques and providing flexibility to add new techniques as they are developed[38].

- Model governance: development or upgrade of model governance systems to support the XAI aspects of MRM (inventory, tiering, documentation, etc.), thus ensuring that the available models meet the required quality, safety and explainability requirements[39].

Beyond this, a holistic approach that encompasses all aspects of the XAI framework is recommended. This includes the use of data analysis tools, the development of APIs for integrating the interpretability and model governance systems described above, the creation of security and auditing mechanisms, and the definition of protocols to ensure compliance with quality and explainability standards.

4. Human factor

A fourth element in embedding XAI into an organization and its processes is the human factor. This includes:

- **Talent recruitment and retention:** develop programs for recruiting and retaining talent specialized in XAI to ensure the availability of professionals with the technical knowledge and experience required to implement XAI in the organization, which is particularly important in a labor market with a shortage of this professional profile.

- **Training:** develop training programs for AI model development teams, model governance teams and AI model users to ensure that everyone involved understands the basic principles of XAI and how to apply them in the specific context of the organization.

- **Culture:** develop a company culture that fosters the implementation of AI model explainability and interpretability. This may include adopting agile methodologies for IA model development, creating a culture of collaboration between model development and model governance teams, and considering explainability as a critical factor in the approval of AI models.

- **Change management:** develop change management programs to ensure the proper adoption of XAI by teams working with AI models in the organization. This includes motivating development teams, analysis of the costs and benefits of explainability, definition of communication protocols with third parties, etc.

In conclusion, AI model explainability and interpretability are key aspects that need to be integrated into an organization and its processes through an appropriate and comprehensive XAI framework, as this is essential to ensure that these models are used in accordance with regulation and best practices.

[38] For this, Management Solutions has ModelCraft™, a proprietary AutoML and component modeling system that incorporates a complete interpretability module. See Management Solutions (2023).

[39] Management Solutions also has Gamma™, a proprietary model governance system that covers all of the above aspects. See Management Solutions (2022).