ManagementSolutions
Making things happen

Explainable artificial intelligence (XAI)

Challenges of model interpretability

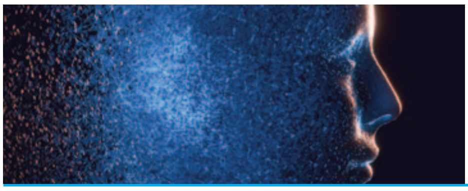
R&D

www.managementsolutions.com

Design and Layout
Marketing and Communication Department
Management Solutions

Photographs
Photographic archive of Management Solutions
iStock
AdobeStock

Contents



Introduction

"Much of what we do with machine learning happens beneath the surface. Though less visible, much of the impact of machine learning will be of this type – quietly but meaningfully improving core operations".

Jeff Bezos[1]

“Artificial intelligence (AI) is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable“[2].

This was the definition of AI offered by John McCarthy, professor at Stanford University, one of the founders of this discipline and co-author of the term “artificial intelligence”.

However, as early as 1950 Alan Turing asked[3]: “can machines think?” and formulated what would later become known as the “Turing test”: a test of a machine's ability to display intelligence indistinguishable from that of a human being. Turing proposed that a human evaluator judge natural language conversations between a person and a machine designed to generate human-like responses. If the evaluator was unable to distinguish the machine from the human, the machine would have passed the test.

Although there is controversy in this regard[4], many authors consider that there are already artificial intelligences that could pass the Turing test, such as GPT-4, from the Open AI Foundation, although GPT-4 itself is not so sure about it (Fig. 1). There are also more sophisticated tests, such as Winograd's schema test, which consists of solving complex anaphora that require knowledge and common sense , something that the current AI does not seem to be able to do yet.

[1] Bezos (b. 1964), J., founder, executive chairman and former CEO of Amazon.

[2] McCarthy (2004). Professor of Computer Science at Stanford University.

[3] Turing (1950). British mathematician, logician, theoretical computer scientist, cryptographer, philosopher and theoretical biologist.

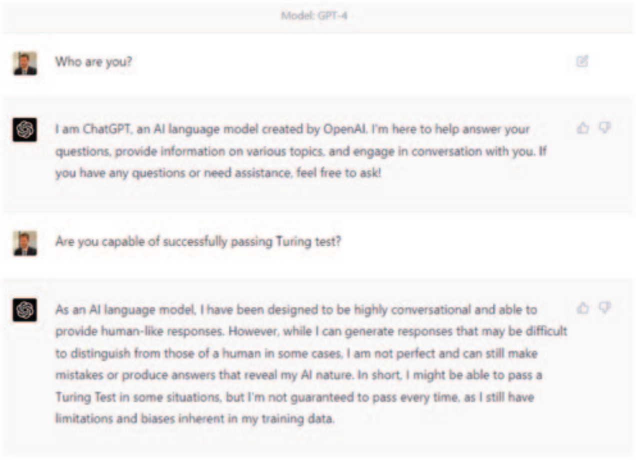
[4] Harnad (2003). Professor of Psychology at the University of Quebec in Montreal (UQAM) and McGill University, and Emeritus Professor of Cognitive Science at the University of Southampton.

[5] A Winograd scheme is a binary choice question where (i) there are two parties mentioned in the question; (ii) pronouns are used to refer to them; (iii) there is ambiguity about who the pronoun refers to; and (iv) there are specific words that can change the correct answer. In an example from Terry Winograd (Professor of Computer Science at Stanford University):

- Question: the city councilmen refused the demonstrators a permit because they [feared/advocated] violence. Who [fears/advocates] violence?
- Answer: [the city councilmen / the demonstrators].

With this, an alternative test to the Turing Test can be generated, using such questions and heavily penalizing wrong answers (see Levesque (2014)).

Figure 1. Conversation with GPT-4 about its ability to pass the Turing test.



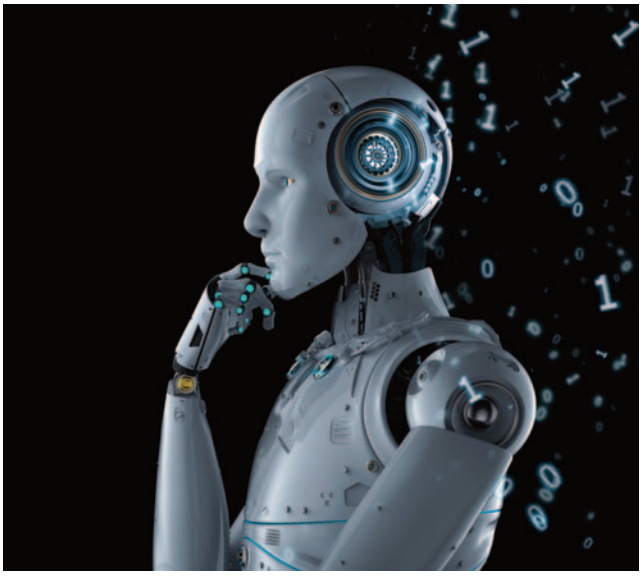
Model: GPT-4

Who are you?

I am ChatGPT, an AI language model created by OpenAI. I'm here to help answer your questions, provide information on various topics, and engage in conversation with you. If you have any questions or need assistance, feel free to ask!

Are you capable of successfully passing Turing test?

As an AI language model, I have been designed to be highly conversational and able to provide human-like responses. However, while I can generate responses that may be difficult to distinguish from those of a human in some cases, I am not perfect and can still make mistakes or produce answers that reveal my AI nature. In short, I might be able to pass a Turing Test in some situations, but I'm not guaranteed to pass every time, as I still have limitations and biases inherent in my training data.



Even so, although the field of AI is not new, dizzying breakthroughs have been made in recent years, with applications ranging from self-driving cars to medical diagnostics, automatic trading, facial recognition, energy management, cybersecurity, robotics or machine translation, to name a few.

A distinguishing feature of today's AI is precisely linked to McCarthy's definition mentioned above: it is not limited to observable methods, and, when it reaches a certain level of complexity, it poses interpretability challenges. In other words: AI models tend to have a high performance, much higher than traditional algorithms; but in each specific case it can be extremely complex to explain why the model has produced a given result.

Although there are applications of AI where it is not as important to be able to understand or explain why the algorithm has returned a particular value, in many cases it is essential and is a regulatory requirement. For example, in the European Union, under the General Data Protection Regulation (GDPR), consumers have what is known as the "right to an explanation"[6]:

> [...] not to be subject to a decision based solely on automated processing [...], such as automatic refusal of an online credit application [...] without any human intervention", and [the data subject] has the right "to obtain an explanation of the decision reached [...] and to challenge the decision".

All this has led to the development of the Explainable Artificial Intelligence (XAI) discipline, which is the field of study that aims to make AI systems understandable to humans[7], as opposed to the notion of "black box", which refers to algorithms in which only the results are observable and the operation of the model is unknown, or the basis for the results cannot be explained.

It can be concluded[8] that an algorithm falls within the XAI discipline if it follows three principles: transparency, interpretability and explainability. Transparency occurs if the processes that calculate the parameters of the models and produce the results can be described and justified. Interpretability describes the ability to understand the model and present how it makes decisions in a human-understandable way. Explainability refers to the ability to decipher why a particular observation has received a particular value. In practice, these three terms are closely linked and are often used interchangeably, in the absence of a consensus on their precise definitions[9].

These principles are achieved through basically two strategies: either develop algorithms that are interpretable and explainable by their nature (including linear regressions, logistic or multinomial models, and certain types of deep neural networks, among others), or use interpretability techniques as tools to achieve compliance with these principles[10].

[6] GDPR (2018), Recital 71.
[7] Vilone et al. (2021). Doctora en Inteligencia Artificial, School of Computer Science, Technological University Dublin.
[8] Roscher et al. (2020). Data Scientist at the Technical University of Munich.
[9] Marcinkevics et al. (2020). Researcher at the Department of Computer Science, ETH Zurich.
[10] iDanae (2022). Chair in Big Data and Analytics (iDanae is a Spanish acronym for intelligence, data, analysis and strategy) created from a collaboration between Management Solutions and the Polytechnic University of Madrid (UPM) in the educational, scientific and technical fields. The Chair aims to promote knowledge creation and dissemination as well as technology transfer, and to foster R&D&I in Data Analytics.

XAI deals both with the techniques to try to explain the behavior of certain opaque models ("black box") and the design of inherently interpretable algorithms ("white box")[11].

XAI is essential for AI development, and therefore for professionals working in this area, due to at least three factors:

- It contributes to building confidence in making decisions that are based on AI models; without this confidence, model users might show resistance to adopting these models.
- It is a regulatory requirement in certain areas (e.g. data protection, consumer protection, equal opportunities in the employee recruitment process, regulation of models in the financial industry).
- It leads to improved and more robust AI models (e.g. by identifying and eliminating bias, understanding the relevant information to produce a certain result, or anticipating potential errors in observations not included in the model's training sample). All of this helps to develop ethical algorithms and allows organizations to focus their efforts on identifying and ensuring the quality of the data that is relevant to the decision process.

Although the development of XAI systems is receiving a great deal of attention from the academic community, industry and regulators, it still poses numerous challenges.

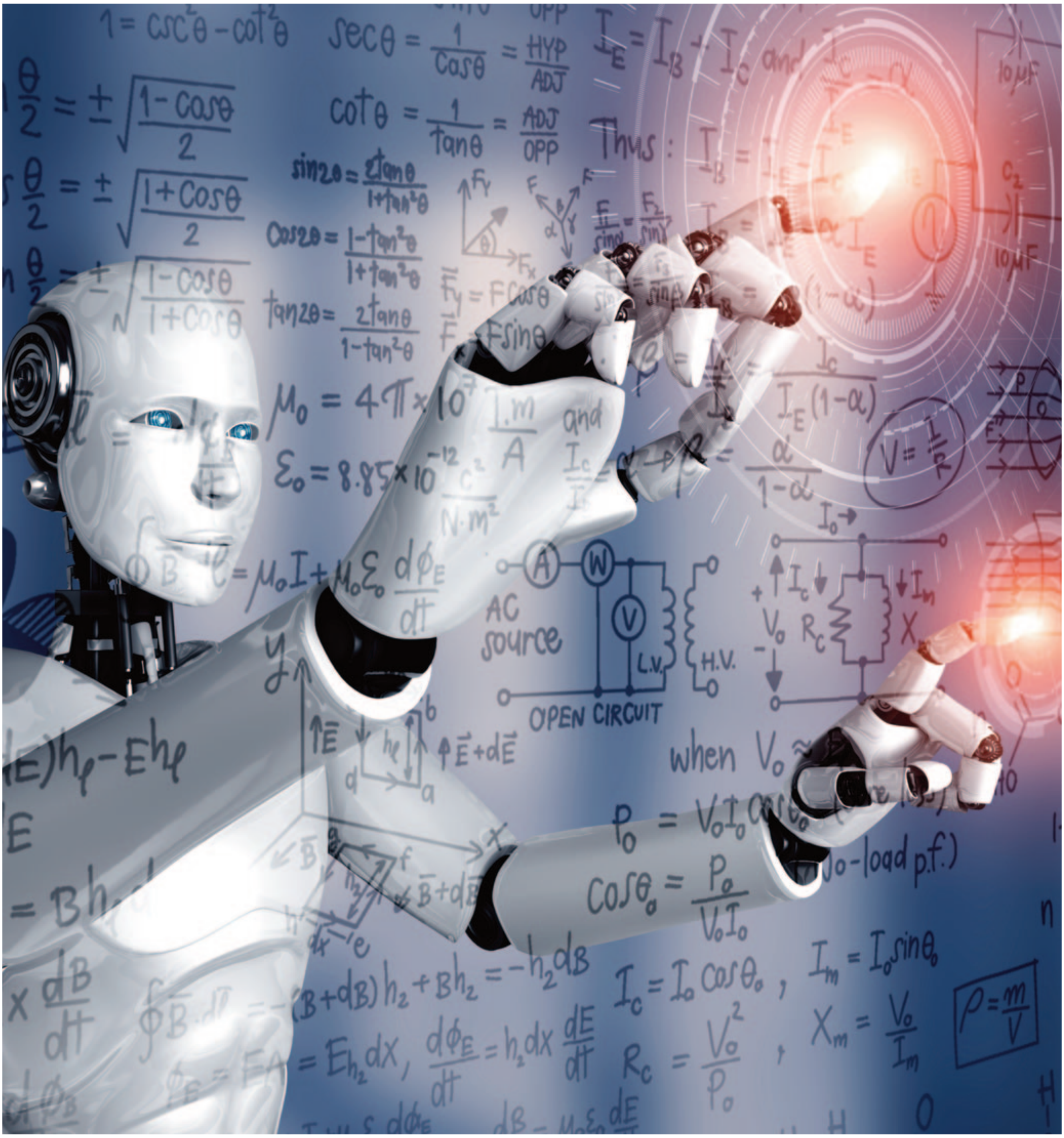
This paper will review the context and rationale for XAI, including XAI regulations and their implications for organizations; the state of the art and key techniques of XAI; and the advances and unsolved challenges in XAI. Finally, a case study on XAI will be provided to help illustrate its practical application.

[11] Sudjianto et al. (2011). Head of Model Risk at Wells Fargo..

Executive summary

"Every technology really needs to be shipped with a special manual – not how to use it but why, when and for what".

Alan Kay[12]



Context and rationale for XAI

1. Digital transformation has enabled access to and exploitation of a vast amount of structured and unstructured data, driving the use of machine learning techniques and artificial intelligence across industries.

2. AI models provide greater predictive power, but they also present risks, such as the presence of undetected bias, lack of understanding of the model, or errors in its application arising from causes such as overfitting, all of which can lead to model distrust. This raises the question of whether it is possible to understand the results of AI algorithms well enough to make appropriate decisions.

3. Explainable Artificial Intelligence (XAI) is a set of processes and methods that enable users to understand and trust the results and products created by machine learning algorithms. This discipline is crucial for an organization to build trust when using AI models, helping to characterize model accuracy, fairness, transparency and understanding of results in AI-based decision making.

4. Academic and business interest in XAI has increased exponentially in recent years, due to this discipline’s ability to address a number of industry concerns regarding the use of AI, such as regulatory requirements, lack of trust, potential misuse, reputational impact, social or human impacts, and other risks.

5. This has led regulators and supervisors in different jurisdictions to establish regulations and guidelines for the appropriate use of AI, including the interpretability aspects of models.

6. In Europe, the European Parliament's General Data Protection Regulation (GDPR) that came into force in 2018 includes a “right to an explanation“ for citizens, requiring companies to be able to explain why an AI model yielded a certain result. This has critical implications for the design and interpretability analysis of AI models.

7. Moreover, in 2021 the European Parliament proposed the Artificial Intelligence Act (AI Act) to regulate the use of artificial intelligence in the European Union. This proposed Regulation sets out a regulatory framework for AI systems, including requirements for ethical development, transparency, security and accuracy, as well as a governance and oversight system. The AI Act classifies AI applications into levels of risk (unacceptable practices, high-risk systems, and low or limited risk systems), and lays down transparency and human oversight requirements for high-risk systems, which will be enforceable across the Union. This is likely to trigger initiatives to adapt to the Regulation, including comprehensive model documentation, interpretability techniques, monitoring dashboards and model alerts.

8. Likewise, in 2019 the European Commission formulated the Ethical Guidelines for Trustworthy Artificial Intelligence, which propose seven key requirements for AI systems to be considered trustworthy: (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. The transparency requirement includes the need for AI models to be explainable. The Guidelines propose evaluation criteria to assess the extent to which an AI model meets these requirements.

9. In the United States, the White House proposed an AI Bill of Rights (AI Bill of Rights) in 2022, pushed by President Joe Biden. This bill sets out five principles or citizen rights regarding AI, including safe and effective systems, protection against discrimination by algorithms, data privacy, notification and explanation, and evaluation and correction by a human in the event of AI failure (fallback). These principles include the explainability of AI models, which requires plain language documentation in addition to

[12] Alan Kay (b. 1940), American Turing Award-winning computer scientist, considered to be the “father of personal computers”.

technically valid, meaningful and useful explanations, and demonstrably clear, timely, understandable and accessible notices of use.

10. The 2019 OECD Principles on Artificial Intelligence promote the use of AI that is trustworthy and respects human rights and democratic values. They were adopted by all 38 OECD member countries and include requirements for transparency and responsible disclosure of AI systems so that those affected by an AI system can understand the outcome.

11. The European Banking Authority's Discussion Paper on Machine Learning for IRB Models, published in 2021, analyzes the relevance of potential barriers to the implementation of machine learning techniques in the IRB approach to capital calculation in financial institutions. The document sets out principles and recommendations to make the use of these techniques compatible with compliance with the European Capital Requirements Regulation (CRR). These recommendations include statistical and economic analysis of the relationship between the input and output variables, documentation that explains the model in a simple way, and the need to detect possible biases in the model.

12. A basic tenet of XAI is the need to embed interpretability and explainability into an organization and its processes. This is done through an XAI framework made up of four elements: interpretability techniques of AI models, integration into model risk management (MRM) processes, IT support and the human factor.

13. Techniques: the core of the XAI framework is based on three main aspects of interpretability: explaining the model design, explaining the model results, and other aspects such as bias detection and periodic model monitoring.

14. MRM: AI model interpretability affects the entire model lifecycle chain, and therefore model risk management. Incorporating the XAI components requires reviewing and updating the organizational and governance framework, the policies and procedures for model development, monitoring, validation, implementation and use, and the audit framework.

15. IT support: to implement an XAI framework, professional IT solutions are needed to support interpretability aspects inherent to AI models, such as model interpretability and governance tools, data analysis systems, APIs, security and auditing mechanisms, and protocols to ensure compliance with quality and explainability standards.

16. Human factor: XAI integration must consider the human factor, including the recruitment and retention of specialized talent, training programs, developing a culture that actively pursues explainable and interpretable AI models, and change management programs to ensure XAI is properly adopted.

17. A fifth additional element central to AI and XAI is data, in that its governance, quality, integrity, consistency, traceability and absence of bias determine the quality of the AI model, and ultimately of the decisions made based on it. However, data issues and their relevance in models are not the subject of this paper, as they have already been extensively covered in previous publications .

Interpretability techniques: state of the art

18. The use of AI techniques has spread to all industries and domains, offering greater predictive power in exchange for greater complexity. This has created the need to explain the results of AI models, which has led to the emergence of increasingly sophisticated techniques for local and global interpretability. These techniques do not completely solve the problem, so other approaches like inherently interpretable models ("white boxes") are being researched to ensure AI model interpretability.

19. The most common approaches to addressing the interpretability issue can be classified into two groups: post-hoc interpretability (global and local interpretability techniques) and inherently interpretable models. There are also complementary strategies, such as model simplification, the use of business-oriented variables, data analysis to identify bias or lack of impartiality, or model development reproducibility analysis.

20. The LIME (Local Interpretable Model-agnostic Explanations) technique can be used to explain a model in a local and agnostic way, meaning that it can provide explanations for a specific prediction without having to understand the underlying model.

21. SHAP (SHapley Additive exPlanations) explains the model locally and globally by evaluating the contribution of each input variable to the model's output.

22. PDPs (Partial Dependence Plots) are used to visualize how a model's output changes when the values of the input variables are changed.

23. White box models are based on algorithms that are inherently interpretable by design. These models are grouped together according to the type of algorithm used, and the parameters to be optimized are usually limited to achieve greater interpretability. This allows for a better understanding of the information and leads to more accurate results, which in turn leads to better decision making, especially in those sectors where interpretability is critical.

[13] See Management Solutions (2020, 2018 and 2015): "Auto machine learning, towards the automation of models", "Machine learning, a key piece in the transformation of business models" and "Data science and the transformation of the financial sector".

24. Despite advances in AI model interpretability, there are still challenges around reproducibility of results, explanation of the most likely predictions sequence, biases in the input data, and fairness and accuracy of explanation. In addition, there is room for improvement in white box models so they can compete in accuracy with black box models in complex problems, as well as in developing new techniques to explain more complex models.

Interpretability use case

25. To demonstrate how the interpretability techniques described above are applied, an illustrative exercise was carried out based on fictitious data generated by IBM and published in Kaggle[14]. The use case seeks to understand the causes that lead employees to leave their jobs, using AI and XAI techniques on the proposed fictitious data.

26. The exercise was conducted with the help of the ModelCraft™[15] component modeling system, which contains multiple relevant AI and XAI techniques, allowing the study to be completed in a much shorter time than usual, and without the need to write code.

27. Different models were trained and validated to explain employee abandonment, among which the random forest yielded the best predictive capacity.

28. To explain the model results, SHAP, LIME and PDP interpretability techniques were used to understand which variables best explain employee attrition, how changes in the most important variables impact different population ranges, and the model's results in individual cases.

29. Proper use and interpretation of the model in this case study would make it possible to anticipate and prevent employee attrition, create profiles with different propensities for attrition, and identify the characteristics of these employees in advance to take appropriate measures. Furthermore, this use case highlights the constraints and difficulties in applying post-hoc interpretability techniques, as well as the fact that using AI models together with an interpretability module can enhance the model's predictive power.

Conclusion

30. Explainable Artificial Intelligence (XAI) is an emerging discipline that seeks to improve the interpretability of AI models by using specific techniques to understand and explain the outcome of these models, and is especially important in highly sensitive domains such as health, security, financial services, and energy.

31. XAI has become a priority for many industries as AI models are growing in complexity and more and more regulation requires their interpretability. A use case developed with ModelCraft™ has demonstrated how these techniques can be employed to understand and explain AI models.

32. In the coming years, it is expected that XAI will continue to develop and grow in importance as AI models become more complex, regulation continues to proliferate, and its use spreads to more highly sensitive domains.

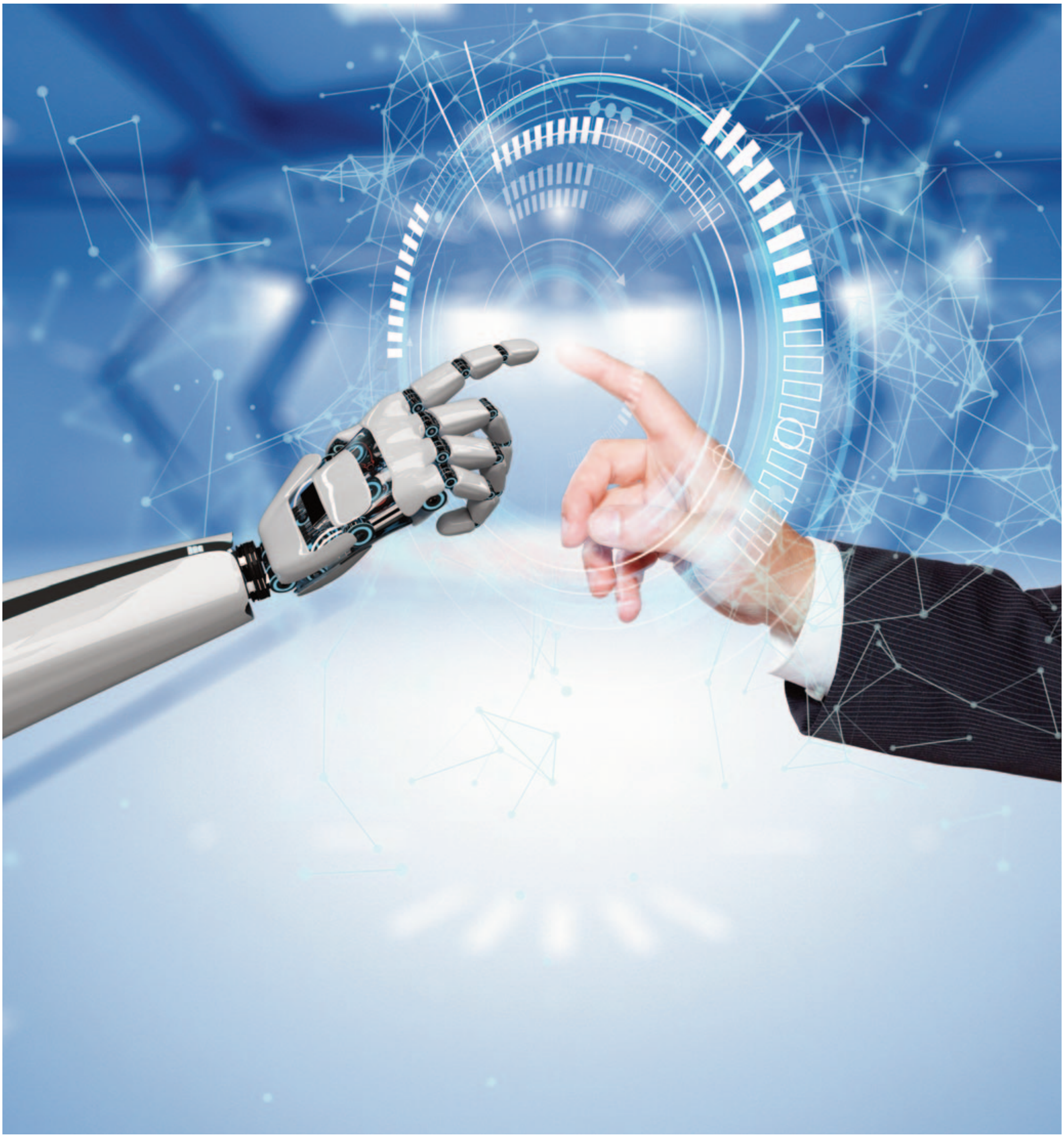
[14] Kaggle (2017). IBM HR Analytics Employee Attrition & Performance.

[15] Management Solutions' proprietary AutoML and component modeling tool. See Management Solutions (2023).

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.

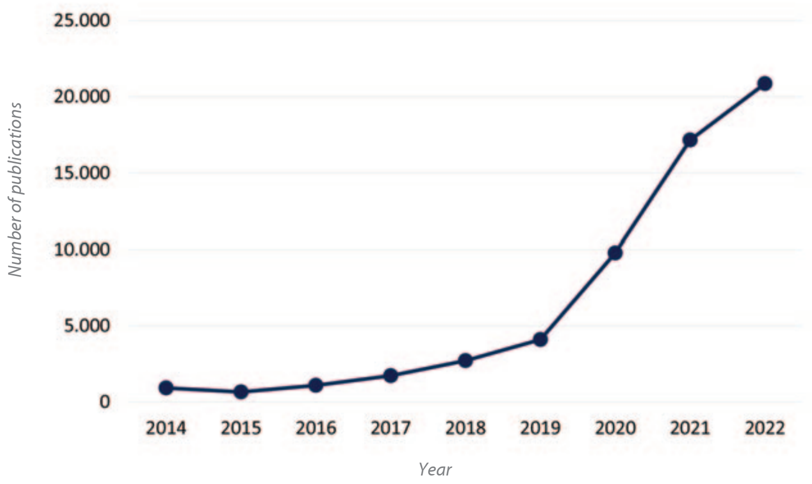
All of this brings up the question of whether it is possible to understand well enough the results that AI algorithms yield, especially when they impact critical decisions, such as medical diagnosis, autonomous driving or fraud detection, among many others.

[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.

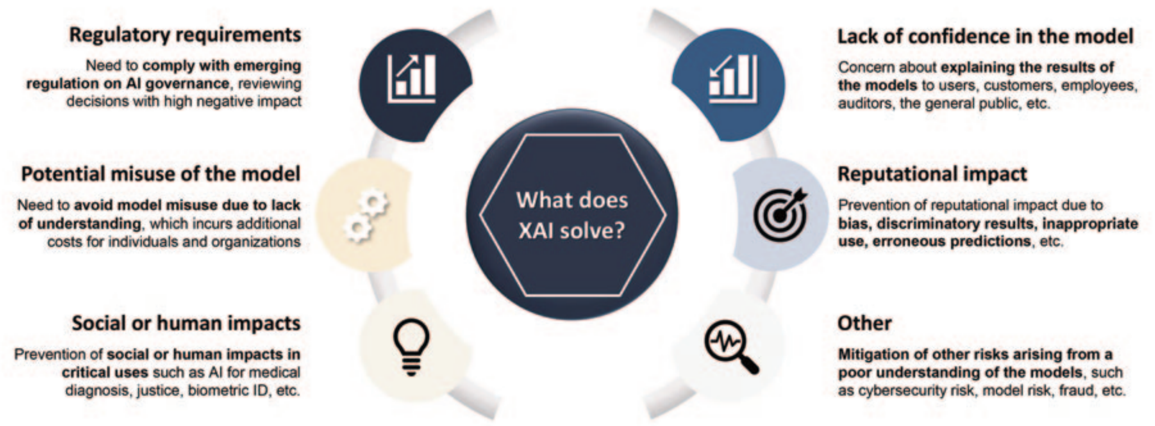
- **Social or human impacts:** the prevention of harmful social or human impacts in critical uses such as AI for the diagnosis of medical diseases, judicial sentences, biometric identification, polygraphs, etc.

- **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


What does XAI solve?
Regulatory requirements
Need to **comply with emerging regulation on AI governance**, reviewing decisions with high negative impact
Potential misuse of the model
Need to **avoid model misuse due to lack of understanding**, which incurs additional costs for individuals and organizations
Social or human impacts
Prevention of **social or human impacts in critical uses** such as AI for medical diagnosis, justice, biometric ID, etc.
Lack of confidence in the model
Concern about **explaining the results of the models** to users, customers, employees, auditors, the general public, etc.
Reputational impact
Prevention of reputational impact due to **bias, discriminatory results, inappropriate use, erroneous predictions**, etc.
Other
Mitigation of other risks arising from a poor understanding of the models, such as cybersecurity risk, model risk, fraud, etc.

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.

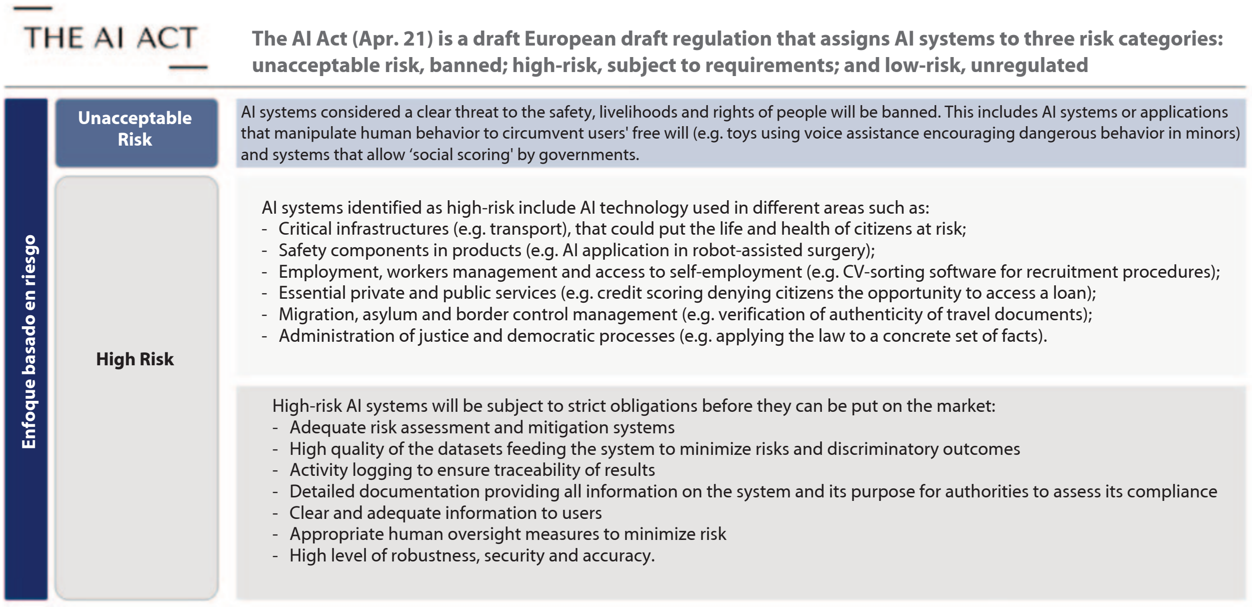
Likewise, what at first sight might appear to be a more relaxed requirement for AI model interpretability, since the Guidelines recognize that some AI models are more difficult to explain, in fact introduces an additional complexity: the need to classify AI models according to their interpretability 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.



THE AI ACT

The AI Act (Apr. 21) is a draft European draft regulation that assigns AI systems to three risk categories: unacceptable risk, banned; high-risk, subject to requirements; and low-risk, unregulated

Enfoque basado en riesgo

Unacceptable Risk

AI systems considered a clear threat to the safety, livelihoods and rights of people will be banned. This includes AI systems or applications that manipulate human behavior to circumvent users' free will (e.g. toys using voice assistance encouraging dangerous behavior in minors) and systems that allow 'social scoring' by governments.

High Risk

AI systems identified as high-risk include AI technology used in different areas such as:
- Critical infrastructures (e.g. transport), that could put the life and health of citizens at risk;
- Safety components in products (e.g. AI application in robot-assisted surgery);
- Employment, workers management and access to self-employment (e.g. CV-sorting software for recruitment procedures);
- Essential private and public services (e.g. credit scoring denying citizens the opportunity to access a loan);
- Migration, asylum and border control management (e.g. verification of authenticity of travel documents);
- Administration of justice and democratic processes (e.g. applying the law to a concrete set of facts).

High-risk AI systems will be subject to strict obligations before they can be put on the market:
- Adequate risk assessment and mitigation systems
- High quality of the datasets feeding the system to minimize risks and discriminatory outcomes
- Activity logging to ensure traceability of results
- Detailed documentation providing all information on the system and its purpose for authorities to assess its compliance
- Clear and adequate information to users
- Appropriate human oversight measures to minimize risk
- High level of robustness, security and accuracy.

- 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.

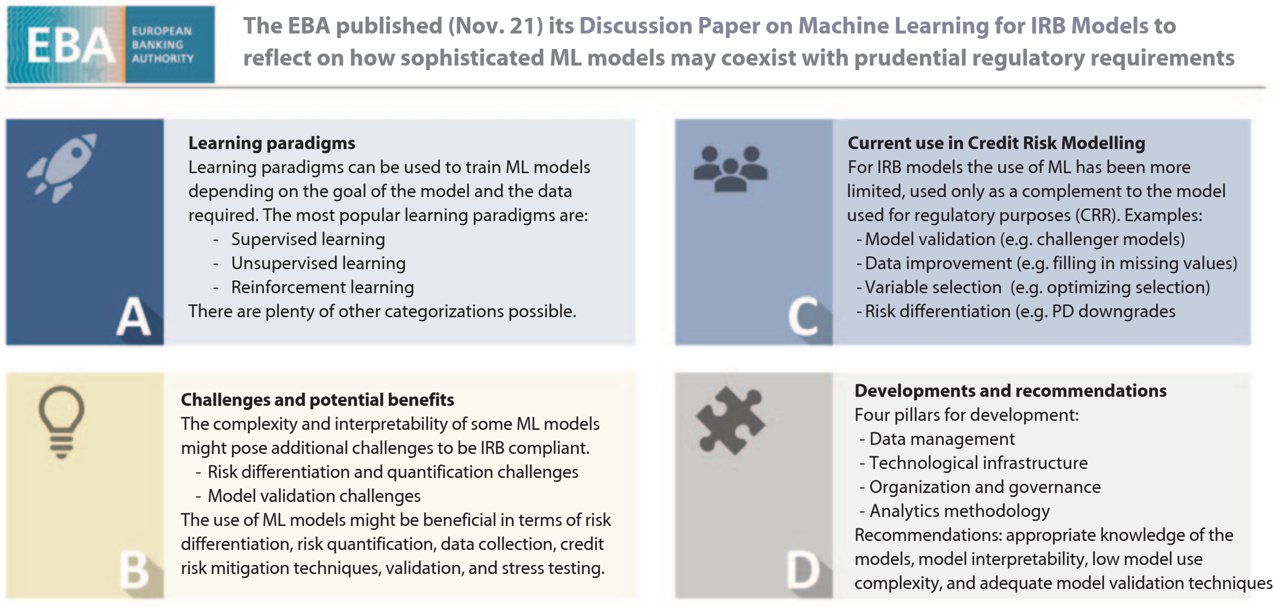
In practice, while the banking industry awaits the final version of the EBA consultation paper, most institutions using machine learning in their IRB models are already adapting their model development, monitoring and validation frameworks to ensure future compliance.

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.



The EBA published (Nov. 21) its Discussion Paper on Machine Learning for IRB Models to reflect on how sophisticated ML models may coexist with prudential regulatory requirements

Learning paradigms
Learning paradigms can be used to train ML models depending on the goal of the model and the data required. The most popular learning paradigms are:
- Supervised learning
- Unsupervised learning
- Reinforcement learning

There are plenty of other categorizations possible.

Current use in Credit Risk Modelling
For IRB models the use of ML has been more limited, used only as a complement to the model used for regulatory purposes (CRR). Examples:
- Model validation (e.g. challenger models)
- Data improvement (e.g. filling in missing values)
- Variable selection (e.g. optimizing selection)
- Risk differentiation (e.g. PD downgrades

Challenges and potential benefits
The complexity and interpretability of some ML models might pose additional challenges to be IRB compliant.
- Risk differentiation and quantification challenges
- Model validation challenges

The use of ML models might be beneficial in terms of risk differentiation, risk quantification, data collection, credit risk mitigation techniques, validation, and stress testing.

Developments and recommendations
Four pillars for development:
- Data management
- Technological infrastructure
- Organization and governance
- Analytics methodology

Recommendations: appropriate knowledge of the models, model interpretability, low model use complexity, and adequate model validation techniques.

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).

Interpretability techniques: state of the art

"By far the greatest danger of artificial intelligence is that people conclude too soon that they understand it".
Eliezer Yudkowsky[40]

Concept

The scientific community[41,42] has proposed numerous definitions of model "interpretability" and "explainability", and tends to make a certain distinction between them, although in practice these concepts are often used interchangeably. Generally speaking, interpretability is linked to the ability to explain to a human being the results of a model (its cause-effect relationship), while explainability is associated with the understanding of an algorithm's internal logic, how it is designed and trained, and the steps followed in decision making to reach a particular result.

Some academic definitions in this regard are:

- Interpretability is the ability to explain or present in terms that are understandable to a human being[43].
- Interpretability is the degree to which a human being can understand the cause of a decision[44].
- The explainability of a model output is the description of how the output of the model was produced[45].
- Explainability is the extent to which the internal mechanics of a machine learning system can be explained in human terms[46].

The need for model explainability and interpretability has favored the emergence of increasingly sophisticated techniques for local and global interpretability of model results, and there is currently some level of standardization and convergence in the use of certain techniques (e.g. PDP, LIME or SHAP).

At the same time, these techniques do not completely solve the problem of interpretability and may yield contradictory or biased results under certain circumstances, which coexists with other factors that may impact model interpretability, such as:

- The reproducibility of results, the model development and implementation process[47], the consistency of the model's predictions and the explanation of the most probable sequence of predictions.
- Potential bias[48] in the input data.
- Fairness[49].
- Accuracy of explanation[50].
- Conceptual soundness of the model[51].

To overcome several of these difficulties, some researchers[52] are developing alternative approaches for improving AI model interpretability, primarily focused on the development of inherently interpretable models ("white boxes").

This section describes the main interpretability techniques, considered standard in the industry, and includes the state of the art on white-box development.

[40] Eliezer Shlomo Yudkowsky (b. 1979), American researcher and writer specializing in decision theory and artificial intelligence, known for popularizing the idea of Friendly Artificial Intelligence and advocating the Singularity.
[41] Gall, R. (2018). Editor at Thoughtworks and The New Stack.
[42] Broniatowsky, D. (2021). Associate Professor, Department of Engineering Management and Systems Engineering, George Washington University.
[43] Doshi-Velez, F., et al. (2017). Professor of Computer Science at the Paulson School of Engineering and Applied Science, Harvard University.
[44] Miller, T. (2019). Lecturer in the School of Computing and Information Systems, University of Melbourne.
[45] Broniatowsky D. (2021).
[46] Gall, R. (2018).
[47] Leventi-Peetz, A.-M., et al. (2022). Scientist of the German Federal Office for Information Security.
[48] Zhou, N., et al. (2021). Senior financial analyst at Wells Fargo.
[49] Ibid.
[50] Jonathon Phillips et al. (2021). Professor of Computer Science and Engineering, National Institute of Standards and Technology (NIST).
[51] Sudjianto, A., et al. (2021).
[52] Ibid.

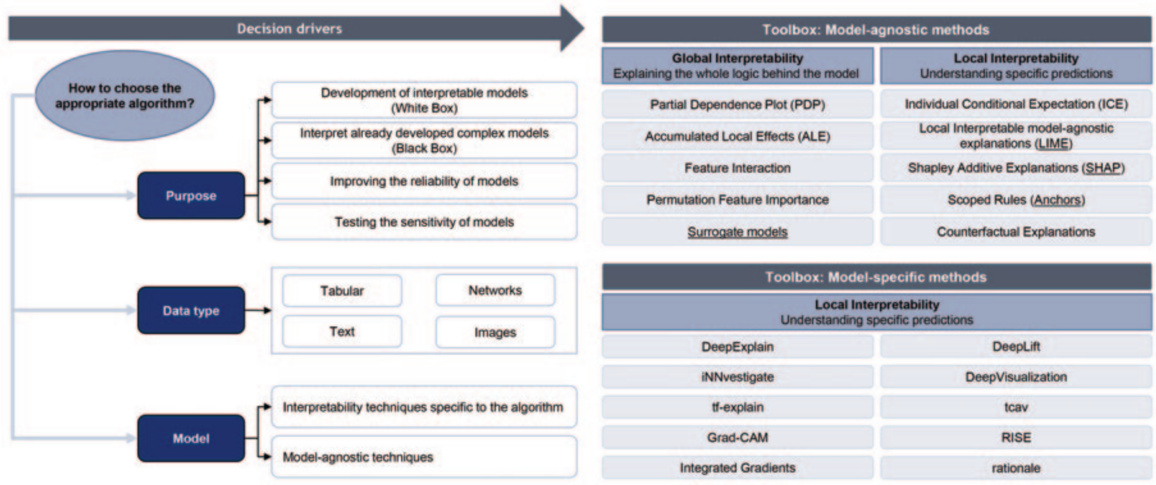

Decision drivers
How to choose the appropriate algorithm?
Purpose
Development of interpretable models (White Box)
Interpret already developed complex models (Black Box)
Improving the reliability of models
Testing the sensitivity of models
Data type
Tabular
Networks
Text
Images
Model
Interpretability techniques specific to the algorithm
Model-agnostic techniques
Toolbox: Model-agnostic methods
Global Interpretability
Explaining the whole logic behind the model
Local Interpretability
Understanding specific predictions
Partial Dependence Plot (PDP)
Individual Conditional Expectation (ICE)
Accumulated Local Effects (ALE)
Local Interpretable model-agnostic explanations (LIME)
Feature Interaction
Shapley Additive Explanations (SHAP)
Permutation Feature Importance
Scoped Rules (Anchors)
Surrogate models
Counterfactual Explanations
Toolbox: Model-specific methods
Local Interpretability
Understanding specific predictions
DeepExplain
DeepLift
iNNvestigate
DeepVisualization
tf-explain
tcav
Grad-CAM
RISE
Integrated Gradients
rationale

Figure 6. Overview of interpretability techniques.

Most common interpretability techniques

The most commonly used interpretability techniques can be grouped according to their approach[53]: post-hoc interpretability and inherently interpretable models. There are also complementary strategies to improve model understanding.

Post-hoc interpretability

Post-hoc interpretability or black box model interpretability techniques focus on explaining the output of already trained models, based on the information provided by the weights assigned to each input variable and the model results. These techniques are useful for understanding model results, although they do not provide information about the training process or explain the internal logic of the algorithm.

They are usually divided into global and local interpretability techniques, in reference to whether the technique explains the entire model as a whole or only the results in a subset of observations or data.

The most common post-hoc interpretability techniques are as follows (for a more comprehensive inventory, see Fig. 6):

- PDP (Partial Dependence Plots). This technique allows visualizing the influence of each individual variable on the model output, excluding the rest of the variables.

- LIME (Local Interpretable Model-agnostic Explanations). This technique allows the explanation of results at the local level, i.e. the explanation of the results of a particular specific observation, based on information from other similar cases.

- SHAP (SHapley Additive exPlanations). This technique allows the local and global explanation of a model's results, that is, the explanation of the influence of each variable on model observations, and the importance of each variable in the model's global results.

- Anchors. This involves the search for decision rules that explain the result.

Inherently interpretable models

Inherent interpretability focuses on the development of "white box" models that are interpretable by design or that can be made interpretable by construction, through a series of conditions dependent on the type of model (e.g. neural networks[54], in particular ReLu[55], and tree-based models[56], among others).

These models allow an explanation of the algorithm's internal logic and the sequence of steps taken to reach a specific result, and therefore allow a better understanding of the results, although their applicability in complex problems may be more limited, depending on the type of algorithm used.

Complementary strategies

Some strategies are used to support model interpretability, such as simplifying the model to facilitate its interpretation, using "business sense" variables, analyzing data to identify biases or lack of fairness in the inputs that may hinder explainability, or analyzing model development or model implementation reproducibility.

[53] iDanae (2022).
[54] Yang, Z., et al. (2019). Department of Statistics and Actuarial Science, University of Hong Kong.
[55] Sudjianto. A., et al. (2011).
[56] Sudjianto. A., et al. (2021).

Post-hoc interpretability

1. PDP

PDP plots[57] *(Partial Dependence Plots) show how an AI model's prediction varies as a function of one or two independent variables in the prediction, i.e. the marginal effect of the predictors. Thus, they make it possible to evaluate the relationship between the independent and dependent variables.*

Synthetically:

- PDPs show the average variation of the prediction graphically on a curve.
- This average variance is obtained by varying a predictor for all the observations in the dataset, and then obtaining the average impact on the prediction.
- A variant of the PDPs are the Individual Conditional Expectation (ICE) graphs, which similarly show how a prediction varies for each specific observation if one of the model's predictors is modified while keeping the rest constant.

2. LIME

LIME[59] (Local Interpretable Model-agnostic Explanations) is a local method that tests how the predictions of a model vary when the input data are perturbed. To do this, LIME applies the following steps:

- Generate synthetic data around an observation in the input data: LIME takes as a starting point a single prediction and the input data that generated it, and generates new input data by perturbing this observation, obtaining the corresponding predictions by the AI model.
- Train a simple model on synthetic data: the resulting dataset composed of the perturbed input data and the predictions generated by the model is used to train a model that is interpretable (e.g. linear models, decision trees).
- Explain the predictions of the simple model as a function of the original data: the importance of each variable in the prediction is obtained - for example, as a function of its coefficients in the regression and its corresponding sign.

Use Case: Loan origination in the banking sector. Use of PDPs

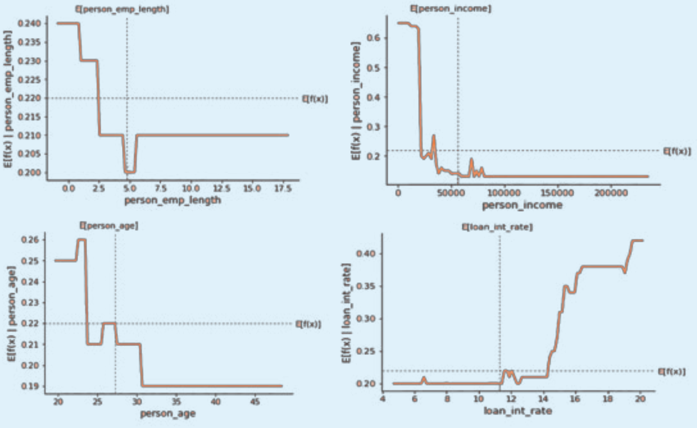
PDPs can be applied to a very common use case in the banking industry: rating customers during the lending process to determine their probability of default. This example uses an anonymized portfolio of mortgage loans with information on their performance in the first three years.

An XGBoost was used, which is a non-additive tree-based model, a feature that may make it difficult to explain. The variables employed by the model during training include the loan amount, its purpose, the borrower's ownership status, years of employment in his current job, and the interest rate, among others.

In this context, a business area may seek to understand why the model assigns a certain probability of default to a certain customer.

A PDP graph shows the explanation that would be obtained at the global level of the variables that have most participated in the result, and that would allow us to see the impact that different ranges of that variable have on the model's prediction (Fig. 7).

Figure 7. PDP for the variables "years employed" (in years), "salary" (annual EUR), "age" (years) and "interest rate" (times one). The X-axis represents the variable under study itself, and the Y-axis represents the impact that different ranges of each variable have on the model's prediction.



[57] Friedman, J. H. (2001). Professor in the Department of Statistics, Stanford University.

[58] Goldstein, A., et al. (2015). Professor in the Department of Statistics, The Wharton School, University of Pennsylvania.

[59] Ribeiro, M. T., et al. (2016). Researcher at Microsoft Research in the Adaptive Systems and Interaction group and Adjunct Professor at the University of Washington.

- Calculate the explainability: the percentage of explainability by LIME is equivalent to the linear model fit coefficient (e.g., R2). It follows that the interpretable model yields a good approximation of the predictions locally.

Formally, an explanation using local subrogated models with LIME can be defined as:

$$Explanation(X) = \arg\min_{g\in G} L(f, g, \pi_X) + \Omega(g)$$

where:

f is a black box model (e.g. a random forest), g is the model that explains f (e.g. a linear regression).

L is the loss function to be minimized in the model (e.g. mean square error), which LIME minimizes.

Ω is the model's complexity (e.g. number of variables selected) decided by the user.

G is the set of possible explanations of the model f.

$\arg$ min represents the value $g \in G$ that minimizes the function $L(f,g,\pi_X) + \Omega(g)$..

π_X represents the amplitude of the perturbations used to generate new observations decided by the user.

3. SHAP

SHAP[60] (SHapley Additive exPlanations) is a model explanation method based on Shapley's Value Theorem , which was proposed in 1952 to distribute the value of a game among the players. SHAP is used to explain the importance of each variable (measured as the average change in the model prediction when the value of the variable varies) in a particular prediction.

Specifically, SHAP uses a combination of baselines, local importance functions and Shapley's Value Theorem to calculate the importance of each variable in an individual prediction.

In this method:

- Shapley values are calculated, where the independent variables are interpreted as players who collaborate to receive the payout.
- The Shapley values correspond to the contribution of each variable to the model prediction.
- The payout is the actual prediction made by the model minus the average value of all predictions.
- Players "split" this payout according to their contribution, and this split is calculated by Shapley's values and reflects the importance of each variable.

This method also makes it possible to obtain interpretations at a global level by calculating the average of the contributions of each variable for each model prediction.

Formally, Shapley values can be defined as the contribution of each variable to the outcome of the model, weighed as a function of all possible combinations of variables used:

$$\phi_j(val) = \sum\nolimits_{S \subseteq \{1,\dots,p\}/\{j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{j\}) - val(S))$$

where val is the prediction of the model for variables included in the set S, with respect to the prediction for variables not included in S:

$$val = \int f(x_1 \dots x_p)\, dP_{x \notin S} - E_X(f(X))$$

where:

x is the vector of variables used in the model.

S is a subset of x.

p is the number of variables used in the model.

$dP_{(x\notin S)}$ represents the set of variables not included in S for which the integration is performed.

E is the expected value of the prediction of X with the f model.

Using these values, SHAP can be used to obtain a local explanation to the model as:

$$Expl(x) = E_X(f(X)) + \sum \phi_j x_j$$

Finally, SHAP is also capable of calculating global explanations through the aggregation of Shapley values in a data set.

[60] Lundberg, S. M., et al. (2017). Research Fellow at the Paul G. Allen School of Computer Science, University of Washington.
[61] Shapley, L. (1953). Professor at the University of California, Los Angeles, in the departments of Mathematics and Economics.
[62] Ribeiro, M. T.; Singh, S.; Guestrin, C. (2018). Researcher at Microsoft Research in the Adaptive Systems and Interaction group and Adjunct Professor at the University of Washington.

4. Anchor

Anchors[62] is a method that explains individual (i.e. local) predictions of black box classification models by finding decision rules called "anchors" that explain the outcome.

- As in LIME, a single prediction and the input data that generated it are taken as a starting point, and new input data are generated by perturbing this observation, obtaining the corresponding predictions by the AI model.
- The local explanation of the prediction is obtained by looking for "if-else" rules that are able to explain the outcome of the model. A rule is considered to explain the prediction if changes in other independent variables not considered in the rule do not modify it.

Formally, an anchor A is defined as:

where:

f is a black box model.

D is an arbitrary distribution used to pertub x.

x is an observation of the dataset to be explained, and z is a sample of D.

Prec is the accuracy of the explanation and **T** is the accuracy required..

One way to find an anchor given any given distribution D is to look for the precision to exceed a threshold with a certain probability (1 - δ), such that:

$$P(Prec\ (A) \geq \tau) \geq 1 - \delta$$

[63] Yang, Z., et al. (2019). Research Fellow in the Department of Statistics and Actuarial Science, University of Hong Kong.

Use case: Loan origination in the banking sector. Use of SHAP

If SHAP is applied on the same case for which a PDP was used, additional local information about a decision of the model is obtained for a given customer.

In this case, using SHAP on a sample of observations results in completely different Shapley values with a variable sign depending on the characteristics of the borrower. Even for clients receiving the same interest rate, the influence of this variable appears to vary due to the greater or lesser importance of the other variables in the model.

However, a "business sense" trend is observed: the higher the interest rate, the more this variable in the model contributes to a higher probability of default. Therefore, using the mean of the Shapley values for each variable to provide an overall interpretation of the model can lead to errors in the explanation if this is understood as a generalization (Fig. 8).

Shapley's values provide an explanation for particular cases such as the following, where it is observed that the probability of default of a client is determined by the mortgage loan conditions, credit history and employment conditions (e.g., salary) (Fig. 9).

Figure 8. Shapley values for the "interest rate" variable in the whole sample versus that variable. The gray bar graph shows the distribution of the variable.

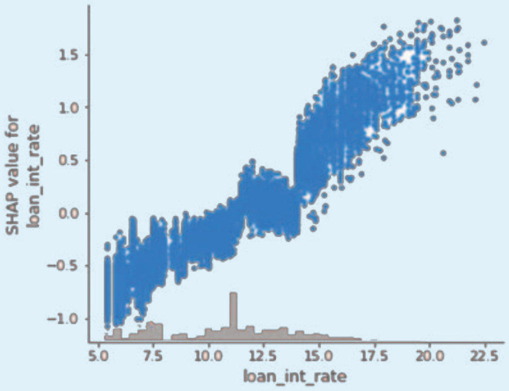
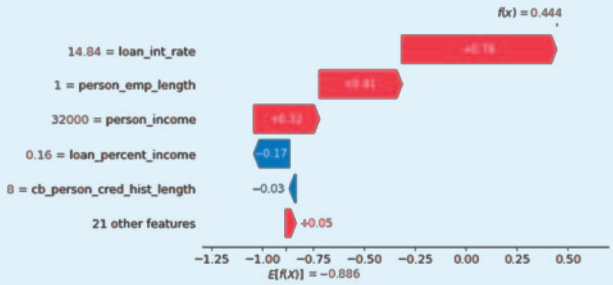
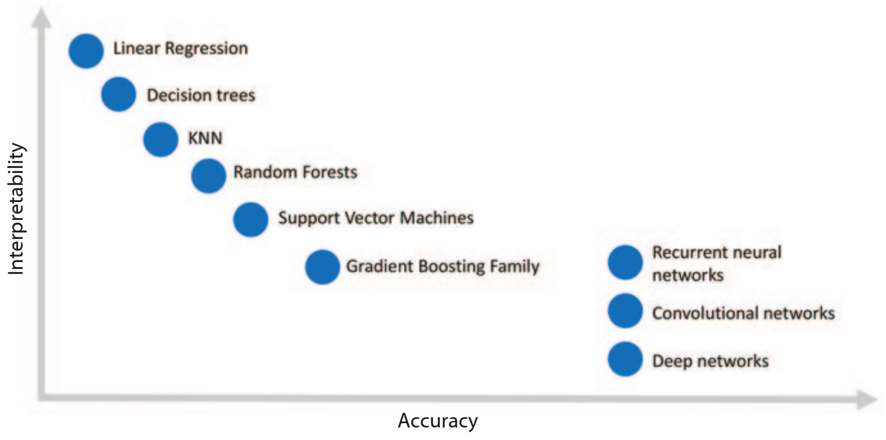


Figure 9. Shapley values influencing the prediction of a client with a denied loan[2].



[1] Scale of the graph shown in log-odds (0 corresponds to a 50% probability).
[2] Log-odds scale graph.

Figure 10. Balance between interpretability and predictive capacity by model families (including white and black boxes).



Development of inherently interpretable models (white box)

Inherently interpretable (white box) models are based on the design of algorithms that, by design, are interpretable and allow the explanation of results at both the global and local levels.

White box models are generally grouped according to the type of algorithm used:

- Linear models, such as linear or logistic regressions.
- Tree-based models, such as decision trees or random trees.
- Rule-based models, such as rule-based systems.
- Deep neural networks, with activation functions such as ReLU or the use of intermediate layers, subject to certain restrictions that make them inherently interpretable[63].

These models are usually developed with constraints on the parameters to be optimized, which allow the models to be interpretable unlike black box models, although they are less accurate (Fig. 10). These constraints include using only "business sense" variables, or restricting:

- The number of variables selected by the model for prediction.
- The number of variables explained by the model.
- The degree of complexity of the decision rules.
- The number of steps in the prediction.
- The depth of the decision trees.
- The length and depth of the neural networks.

Inherently interpretable models provide more accurate results, as they allow for a better understanding of the information, which in turn leads to better decision making. This is especially necessary in those sectors where interpretability is a critical factor in final decisions.

Two aspects relevant to the construction of inherently interpretable models are detailed below: the concept and development of interpretable supervised and unsupervised learning, and the application of other factors in the interpretability domain.

1. Interpretable supervised and unsupervised learning

Although current research is moving towards the development of inherently interpretable models, there is no mathematical formalism that fully describes the construction of these models under whatever initial conditions and algorithms used.

The state of the art is the construction of these models under initial conditions that make them more easily interpretable or equivalent to other interpretable models. One of the ways to define this interpretability condition in model training is to modify the loss[66] function that is minimized during its training, including a penalty for low interpretability, which depends on an imposed model interpretability condition f:

$$Min\left(\frac{1}{n}\sum Loss(f, z_i) + C \cdot InterpretabilityPenalty(f)\right)$$

[64] Rudin, C., et al. (2022). Professor of Computer Science, ECE, Statistics and Biostatistics and Bioinformatics at Duke University.

For example, sparsity is one of the conditions used in model development to qualify a model as more explainable with respect to the rest. This condition can be added to the loss function as:

$$Min\left(\frac{1}{n}\sum Loss(f, z_i) + \varphi(f)\right)$$

such that φ(f) is a regularization function that penalizes the loss being proportional to the sparsity of the model (e.g. if the sparsity is reduced, that term of the loss function will also be reduced).

Some authors[67] have formalized the creation of inherently interpretable models for certain families as: models based on decision trees (e.g. SIMTree or single-index model tree, which generates a single-index model tree for each terminal node), or the simplification of networks with ReLu activation function, which are shown to be equivalent to a set of local linear models.

Other authors[68] have focused on defining the characteristics that inherently interpretable models should meet, in order to optimize them during the modeling process, such as:

- Additivity of the input variables, so that their effects are aggregated in the model in a simple way.
- Sparsity, and the optimization of models to meet this condition.
- Linearity of input variables versus model output.
- Monotonicity, so that the relationship between the input variable and the outcome to be predicted is monotonic for as many ranges as possible.
- Decoupling of concepts during the neural network training, which refers to maintaining as much as possible the information about a given concept in specific network paths (i.e. in the face of information about the same concept passing through a greater number of neurons and paths dispersed in the network).
- Dimensionality reduction as a visual tool to facilitate post-hoc explanations to humans.

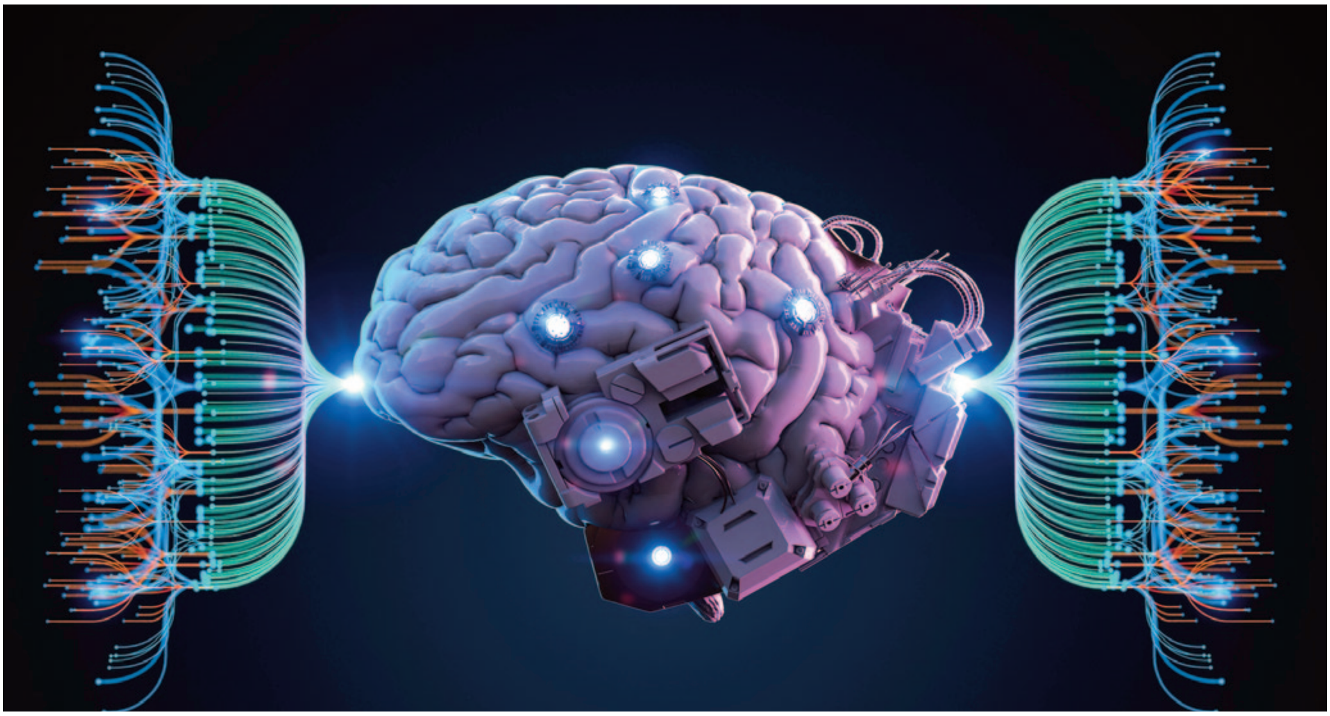
2. Other impact factors

In combination with the challenges shown in this section, there are additional key elements that can be considered to improve model interpretability, such as model fairness, absence of bias in the input data, potential expert components, or adequate performance and model control framework to avoid errors in model interpretation.

Because of their relevance, as indicated above[69], these elements have also been highlighted in the AI Act as essential requirements for high-risk AI systems.

Nowadays, there are multiple techniques and methods to evaluate model performance, and to prevent overfitting issues. There are also several ways to evaluate the error produced by models and the balance between bias and variance error.

[65] Sudjianto. A., et al. (2021).
[66] Rudin, C., et al. (2022).
[67] See section on regulation.



However, due to constraints on the use of personal data introduced by data protection regulations, one of the greatest complexities at the moment is in detecting and correcting potential biases (e.g. due to race, gender, religion, political or sexual orientation, beliefs or social position) in AI models, especially when the variables have not been stored and are therefore not available for analysis.

In this regard, several techniques for identifying unbiased input variables have been proposed by academia, such as:

- Interpretability analysis through Causal Bayesian Networks[68] as a quantification of the degree of model fairness.
- Definition[69] of fairness metrics, such as demographic parity, predictive ratio parity, false positives and equal false negatives in segments susceptible to bias.

Among these metrics, counterfactual fairness provides a measure of how similar the results of a model are to individuals (observations) with the same characteristics, but with slightly different bias-sensitive attributes.

Advantages and disadvantages of the most common interpretability techniques

As a general rule, there is no interpretability technique that can provide a single, global and intuitive explanation for any scenario. Interpretability techniques are usually combined under various use cases and scenarios to verify that they provide reproducible explanations applicable to different groups of observations.

When selecting which of these techniques to use, it is advisable to consider the advantages or disadvantages of their implementation (Fig. 11).

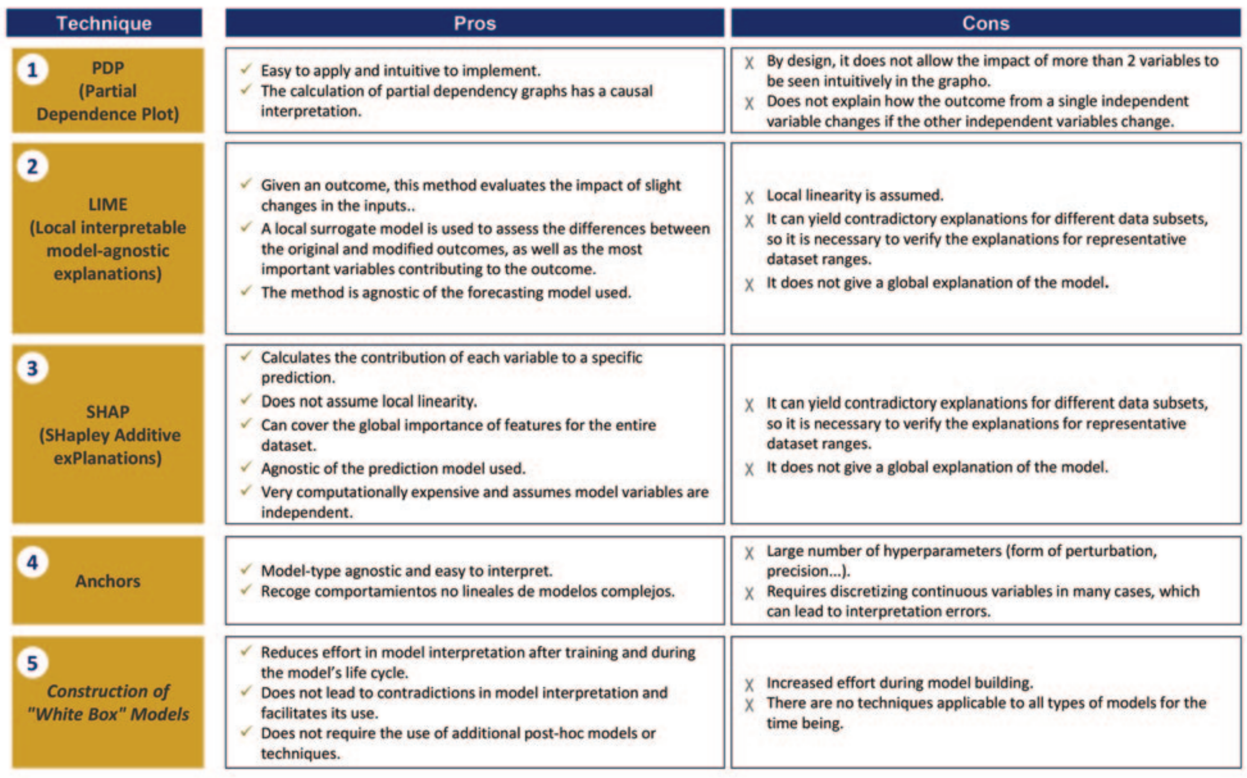
Latest trends and challenges

Despite advances in model interpretability, there are still challenges in explaining the results (Fig. 12).

First, model interpretability is still constrained by a number of factors such as the reproducibility of the results[70], the model training and implementation process, the consistency of model predictions, the explanation of the sequence of most likely predictions, the biases in the input data, as well as the fairness and accuracy of the explanation.

Secondly, currently available XAI techniques only allow either local explanations (i.e. for a single observation or data) or global explanations (i.e. for the whole data set). This means there is a need to develop techniques that allow midrange explanations, i.e. explaining results for groups or subsets of data in a consistent manner . In addition, without an in-depth analysis,

Figure 11. Comparison of the most common interpretability techniques.



Technique	Pros	Cons
1 PDP (Partial Dependence Plot)	✓ Easy to apply and intuitive to implement. ✓ The calculation of partial dependency graphs has a causal interpretation.	X By design, it does not allow the impact of more than 2 variables to be seen intuitively in the grapho. X Does not explain how the outcome from a single independent variable changes if the other independent variables change.
2 LIME (Local interpretable model-agnostic explanations)	✓ Given an outcome, this method evaluates the impact of slight changes in the inputs.. ✓ A local surrogate model is used to assess the differences between the original and modified outcomes, as well as the most important variables contributing to the outcome. ✓ The method is agnostic of the forecasting model used.	X Local linearity is assumed. X It can yield contradictory explanations for different data subsets, so it is necessary to verify the explanations for representative dataset ranges. X It does not give a global explanation of the model.
3 SHAP (SHapley Additive exPlanations)	✓ Calculates the contribution of each variable to a specific prediction. ✓ Does not assume local linearity. ✓ Can cover the global importance of features for the entire dataset. ✓ Agnostic of the prediction model used. ✓ Very computationally expensive and assumes model variables are independent.	X It can yield contradictory explanations for different data subsets, so it is necessary to verify the explanations for representative dataset ranges. X It does not give a global explanation of the model.
4 Anchors	✓ Model-type agnostic and easy to interpret. ✓ Recoge comportamientos no lineales de modelos complejos.	X Large number of hyperparameters (form of perturbation, precision...). X Requires discretizing continuous variables in many cases, which can lead to interpretation errors.
5 *Construction of "White Box" Models*	✓ Reduces effort in model interpretation after training and during the model's life cycle. ✓ Does not lead to contradictions in model interpretation and facilitates its use. ✓ Does not require the use of additional post-hoc models or techniques.	X Increased effort during model building. X There are no techniques applicable to all types of models for the time being.

[68] Oneto, L.,Chiappa, S., (2020)
[69] Zhou, N., et al. (2021). Senior financial analyst at Wells Fargo.
[70] Leventi-Peetz, A.-M., et al. (2022).
[71] While SHAP is able to obtain explanations for subsets through weighted averages of Shapley values, these explanations may vary depending on the granularity of the subset data.

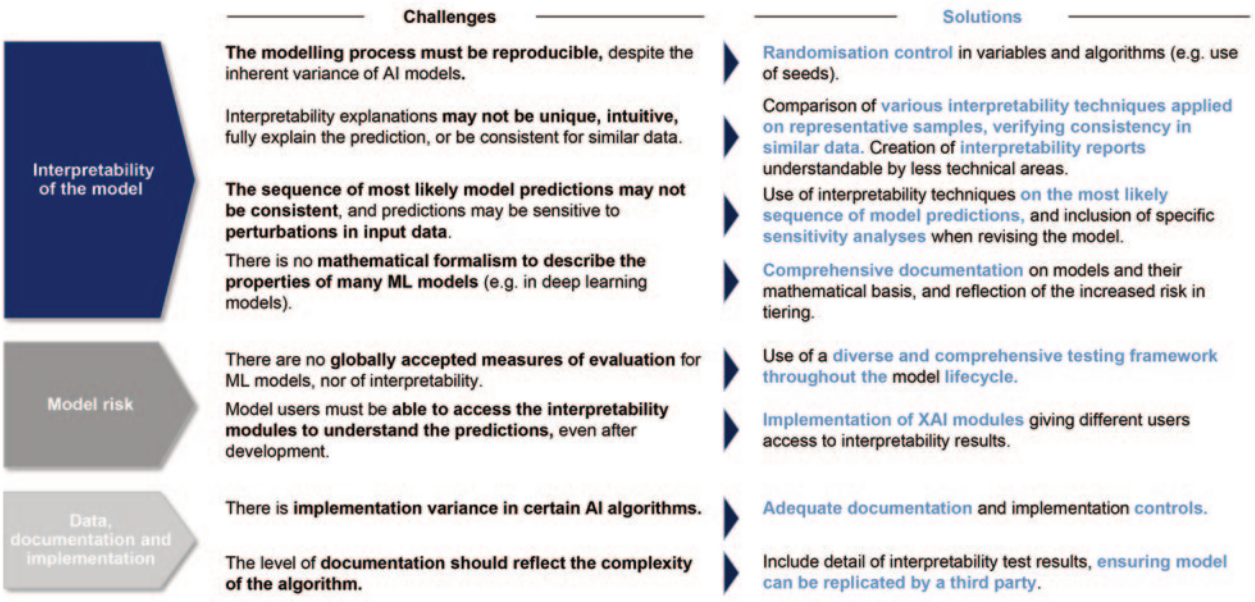
the results yielded by different interpretability techniques at different levels may initially appear contradictory (e.g. if "average" global results are compared with local results in a particular environment).

Thirdly, improvements are still needed in the development of white box models, since, despite the progress made in recent years, these models are still not able to compete in accuracy with black box models in complex problems.

Finally, the need to explain more complex models (e.g. certain types of deep neural networks) remains an unresolved challenge.

In this regard, new techniques are being developed to improve the interpretability of the models, such as the use of information from the intermediate layers of deep neural networks, the aggregation of interpretability metrics to measure the explainability of the models, the development of adversarial models to quantify the degree of explainability, the limitation of the parameters to be optimized to increase their interpretability, or the use of visualization techniques to facilitate the understanding of the results.

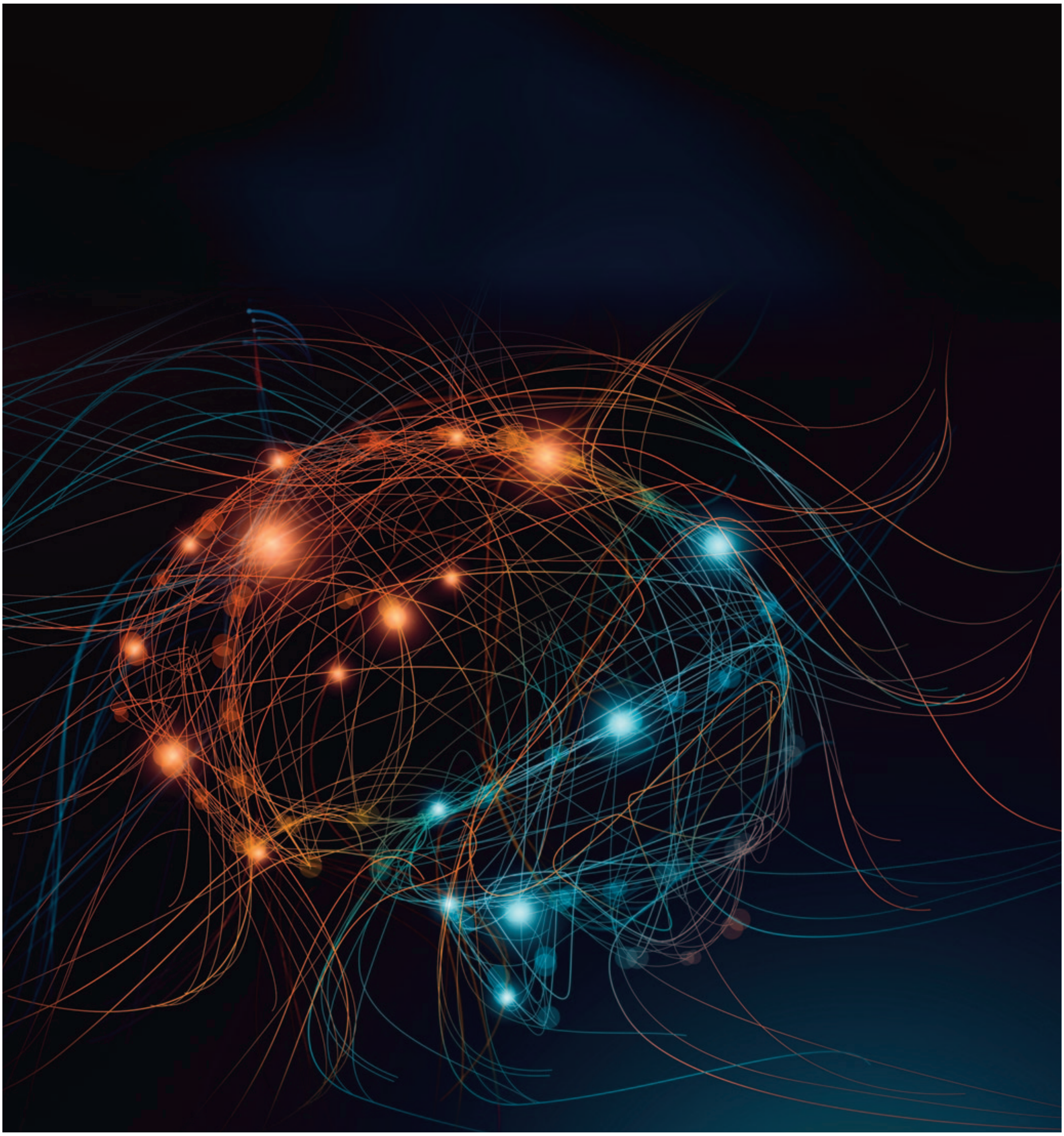
Figure 12. Common challenges in the interpretability of AI models.



	Challenges	Solutions
Interpretability of the model	**The modelling process must be reproducible,** despite the inherent variance of AI models.	Randomisation control in variables and algorithms (e.g. use of seeds).
	Interpretability explanations **may not be unique, intuitive,** fully explain the prediction, or be consistent for similar data.	Comparison of various interpretability techniques applied on representative samples, verifying consistency in similar data. Creation of interpretability reports understandable by less technical areas.
	The sequence of most likely model predictions may not be consistent, and predictions may be sensitive to **perturbations in input data**.	Use of interpretability techniques on the most likely sequence of model predictions, and inclusion of specific sensitivity analyses when revising the model.
	There is no **mathematical formalism to describe the properties of many ML models** (e.g. in deep learning models).	Comprehensive documentation on models and their mathematical basis, and reflection of the increased risk in tiering.
Model risk	There are no **globally accepted measures of evaluation** for ML models, nor of interpretability.	Use of a diverse and comprehensive testing framework throughout the model lifecycle.
	Model users must be **able to access the interpretability modules to understand the predictions,** even after development.	Implementation of XAI modules giving different users access to interpretability results.
Data, documentation and implementation	There is **implementation variance in certain AI algorithms.**	Adequate documentation and implementation controls.
	The level of **documentation should reflect the complexity of the algorithm.**	Include detail of interpretability test results, ensuring model can be replicated by a third party.

Interpretability use case

"Fools ignore complexity. Pragmatists suffer from it.
Some can avoid it. Geniuses remove it"
Alan Perlis[72]



Approach

This section presents a use case for AI interpretability to illustrate how the XAI techniques described in the previous section are applied.

The selected use case addresses the problem of employee retention in an organization, focusing on understanding and explaining the causes that lead employees to leave their jobs. Identifying these factors can enable organizations to take preventive measures and develop strategies to improve job satisfaction and talent retention.

This use case is based on a fictitious dataset generated by IBM and published in Kaggle[73]. This dataset contains information about an organization's employees, including demographic characteristics, data about their job title, and whether they have left the company.

In the year under review, the company has an employee attrition rate of 16%, 6% above the historical average, and is concerned about finding out the causes to be able to develop a remediation plan.

The main variables present in the data set include:

- Level of education (from “high school“ to “Ph.D.”).
- Satisfaction with the work environment (from “low“ to “very high”).
- Job involvement (from “low” to “very high”).
- Job satisfaction (from “low“ to “very high”).
- Performance rating (from “low” to “outstanding”).
- Satisfaction with labor relations (from “low“ to “very high”).
- Work/life balance (from “bad” to “optimal”).
- Years since last job promotion (numerical variable).
- Monthly salary (numerical variable).
- Years in current job (numerical variable).
- Distance from home to work (numerical variable).
- Number of companies in which the employee has worked (numerical variable).
- Role in current job (categorical variable, includes “Manager”, “Director”, “Research Scientist”, among others).

The focus of this use case was to train and validate different artificial intelligence models to predict employee attrition, using XAI techniques to analyze and understand the behavior and decisions of the selected models.

To simplify and streamline the process, the ModelCraft™ component modeling system, which contains multiple relevant AI and XAI techniques, was used. This system allowed the study to be carried out efficiently and without the need to write code.

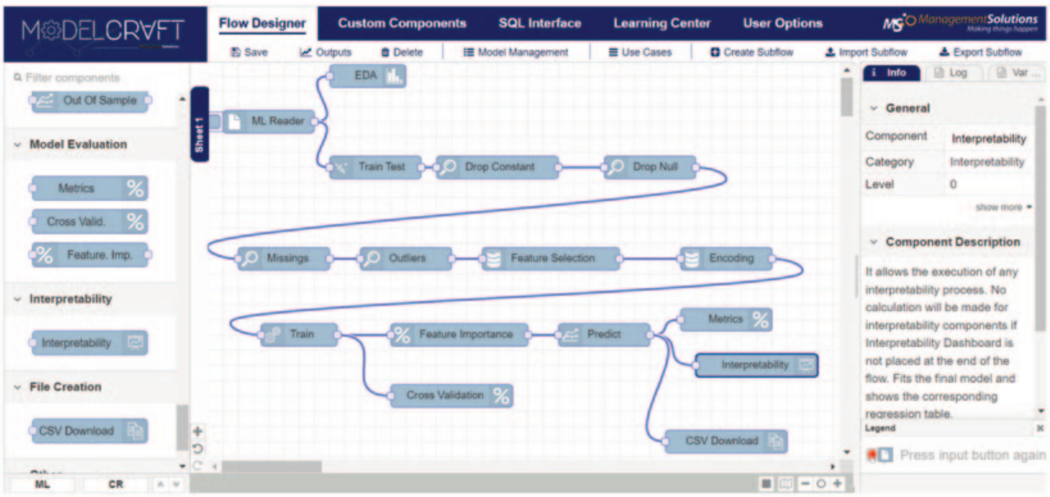
Throughout the use case, the SHAP, LIME and PDP interpretability techniques were applied to analyze the selected models and understand which variables influence employees' decisions to leave their jobs. In addition, this use case explored how these variables interacted with each other and how they affected different segments of the employee population.

At the end of the use case, the effectiveness and limitations of the interpretability techniques used will be evaluated. There will be a discussion on how the combination of artificial intelligence models and interpretability modules can improve the predictive capability and understanding of the models, thus facilitating data-driven decision making in the business domain.

[72] Alan Jay Perlis (1922-1990), american computer scientist, PhD in Computer Science from MIT and professor at Purdue University, Carnegie Mellon University and the University of California at Berkeley, known for his pioneering work in programming languages and for being the first winner of the Turing Award.

[73] Kaggle (2017). IBM HR Analytics Employee Attrition & Performance.

Figure 13. Modeling flow in ModelCraft™.



Component Interpretability
Category Interpretability
Level 0

It allows the execution of any interpretability process. No calculation will be made for interpretability components if Interpretability Dashboard is not placed at the end of the flow. Fits the final model and shows the corresponding regression table.

Modeling process

The modeling process is carried out in three phases: data engineering, modeling and model interpretability analysis.

1. Data engineering

Data engineering is the initial phase in which the data set is prepared and processed for use in the creation of artificial intelligence models. In this case, the following actions were performed:

- Definition of the scope of application: in this case, the population was taken as all employees who had been on sick leave in the previous two years.

- Data cleansing: data quality was verified and records with missing or inconsistent information were eliminated or corrected.

- Variable transformation: categorical variables were converted into numerical variables using techniques such as one-hot encoding or ordinal encoding. In addition, numerical variables were normalized or standardized when necessary.

- Variable selection: the most relevant variables for predicting employee attrition were identified using variable selection techniques such as Pearson correlation, feature importance in tree-based models or recursive feature elimination.

— Feature engineering: new variables were generated from existing ones to analyze whether they were better predictors of employee turnover, such as “total satisfaction”, which was constructed as the sum of the scores of the variables “Satisfaction with the work environment”, “Job satisfaction”, “Performance rating”, “Work-life balance”, “Job involvement “ and “Satisfaction with labor relations”.

- Train-test split: the dataset was divided into two subsets: training and testing. The training subset was used to tune and optimize the artificial intelligence models, while the test subset was used to evaluate the performance and predictive power of the models.

2. Model development

In this phase, different artificial intelligence models were trained and validated using the training subset. Specifically, several of the most common machine learning algorithms and traditional models, such as logistic regression, decision trees, support vector machines, neural networks and random forest, were fitted and compared to select the model with the best performance.

To avoid overfitting and to optimize the hyperparameters of the models, cross-validation and grid or random search techniques were used. In addition, model complexity was given particular consideration when selecting a specific algorithm during training in order to facilitate model interpretation.

For this purpose, a model development flow was generated in ModelCraft™ (Fig. 13).

To select the model with the best predictive power, its performance on the test subset was evaluated using metrics such as accuracy, sensitivity, specificity and area under the ROC curve (AUC-ROC). These metrics allowed us to evaluate the effectiveness of the selected model in terms of its ability to correctly predict employee attrition on previously unseen data.

All things considered, the random forest yields superior performance results, although it poses an interpretability challenge in understanding its predictions. This model has considered 300 decision trees and has yielded an accuracy of 75% and a sensitivity of 84%. Therefore, these are very reliable predictions and false negatives are rarely obtained. This is relevant for this use case, where the company would foreseeably want to reduce this type of error as much as possible.

3. Interpretability analysis

In this last phase, interpretability techniques were applied to analyze and understand the behavior and decisions of the selected model. Specifically, the objectives of the analysis were:

- To understand which variables were most important in decision making for the organization at a global level, for which purpose a comparison of the importance of each variable was used.
- To understand how changes in the most important variables impact different population ranges.
- To understand model results in specific cases where a certain probability of abandonment is observed.

In this use case, SHAP, LIME and PDP techniques were used to explain how the model made decisions and how the inputs influenced the predictions.

SHAP allowed to obtain local and global interpretability results, which provided an interpretation of the importance of each variable, and LIME allowed us to perform an intuitive analysis of local interpretability that made it possible for us to explain the outcome of the model for each employee based on simpler linear models. As a complement, PDP graphs allowed visualization of how changes in each variable impacted the model's prediction.

Thus, the following distribution of the importance of each variable was obtained (Fig. 14).

In this case, it was observed that the variable with the greatest importance in predicting employee attrition (15.65%) was "overall satisfaction", a synthetic indicator defined as a weighted average of six elements (work environment, suitability of functions and areas to the position, internal rating, work/life balance, relationship with colleagues and supervisors, and employee position and responsibility).

This result was intuitive and showed that the "overall satisfaction" variable was well designed. However, the next three variables by importance (length of service, salary, and distance from home to work) appeared to have a high influence on employee turnover, which collectively doubled that of the "overall satisfaction" indicator.

To understand how each variable was influenced individually, the PDPs were studied (Fig. 15).

In terms of length of service, the trend was reversed after three years: employees with intermediate length of service were, on average, the least likely to leave the company. For overall satisfaction, an intuitive trend was observed: higher satisfaction reported in internal surveys resulted in a lower quit rate.

To complement the previous analysis, LIME was used for a case-by-case analysis of the values of variables influencing the likelihood of certain employees leaving the company. Fig. 16 shows two employees with different quit probabilities obtained using the model. LIME shows an explainability metric representing how good a linear fit it has obtained using the local surrogate model to explain these predictions.

It is interesting to see how the most significant causes of employee abandonment in these two cases do not necessarily correspond to the most influential variables at the global level. While overall satisfaction appears to contribute to explaining the likelihood of employee abandonment in case 1, it does not seem to have a significant impact in case 2, where the

Figure 14. Global interpretability of the random forest model using SHAP, where the Shapley values are used to obtain the importance of the variables.

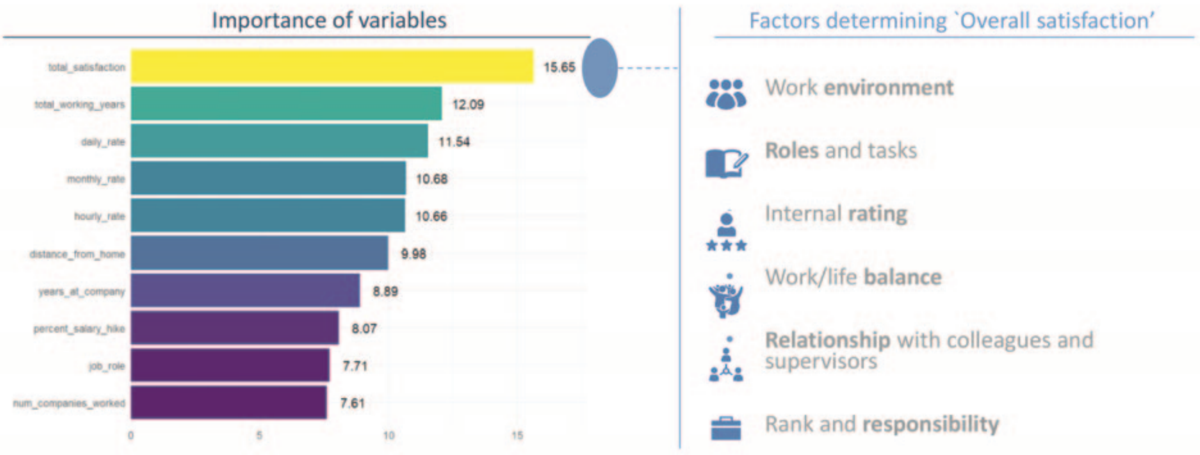
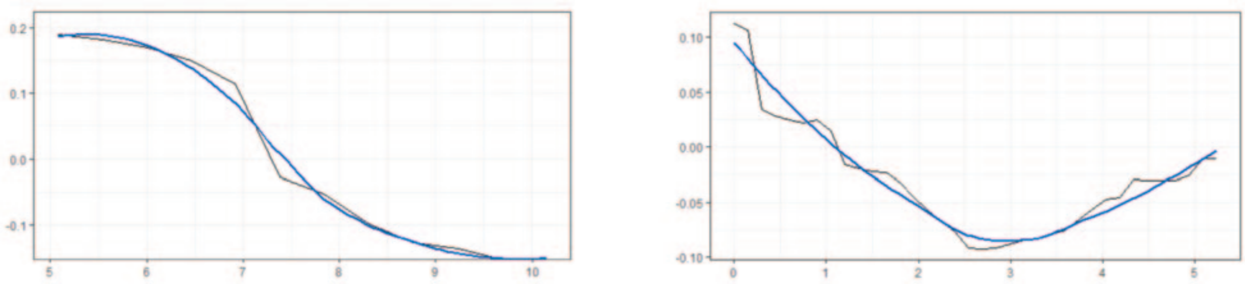


Figure 15. PDP plots for the variables "total satisfaction" and "length of service".



probability of leaving is higher.

This reflects the difficulties in interpreting this model, which can be generalized to similar models: although overall satisfaction can explain the average probability of employee abandonment, this conclusion is a generalization, as there are individual and group cases in which employee abandonment is explained to a greater extent by other variables.

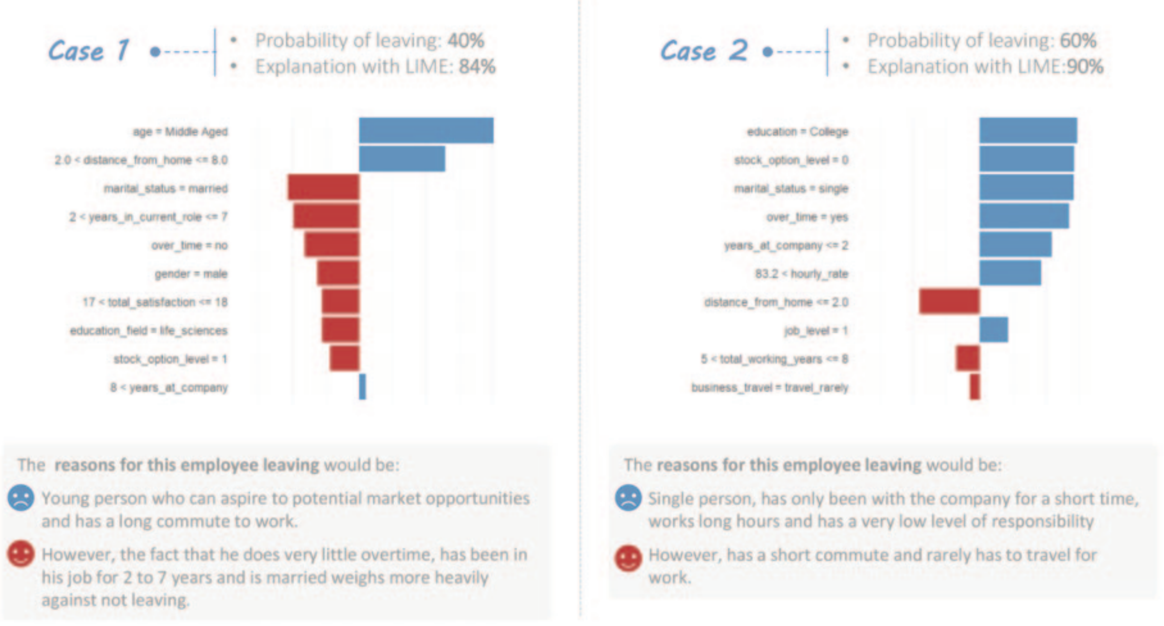
Conclusions of the use case

Several conclusions and lessons learned can be drawn from this artificial intelligence interpretability use case that may be useful in future uses of AI and XAI models:

- **Model use:** the correct use and interpretation of the model in this case can make it possible to anticipate and prevent employee turnover. Among the uses that can be made of the model is the ability to create different profiles with a propensity to leave and identify the characteristics of these employees in advance to take appropriate measures, which in the long term can contribute to reducing the level of turnover in the organization.

- **Model selection:** the modeling process demonstrated the importance of comparing and validating different machine learning algorithms to select the model with the best predictive capability. In this case, the random forest model proved to be the most suitable for predicting employee attrition.

Figure 16. Local interpretability of the random forest model using LIME.



The reasons for this employee leaving would be:

☹ Young person who can aspire to potential market opportunities and has a long commute to work.

☺ However, the fact that he does very little overtime, has been in his job for 2 to 7 years and is married weighs more heavily against not leaving.

The reasons for this employee leaving would be:

☹ Single person, has only been with the company for a short time, works long hours and has a very low level of responsibility

☺ However, has a short commute and rarely has to travel for work.

- **Importance of interpretability:** the use of interpretability techniques, such as SHAP, LIME and PDP, provided a deeper understanding of how the model makes decisions and how inputs influence predictions. This information is crucial to validate the applicability of the model in a real-world context and to ensure that predictions are based on relevant and meaningful features.

- **Influential variables:** the interpretability analysis allowed us to identify the most relevant variables for predicting employee attrition. These variables can be useful in developing retention strategies and improving job satisfaction. In addition, understanding how these variables interact with each other and how they affect different segments of the employee population can enrich the analysis and facilitate data-driven decision making.

- **Practical implementation:** the use case demonstrates the feasibility and usefulness of applying AI and XAI techniques in a realistic scenario, using fictitious data but representative of a business situation. This approach can be adapted to other business contexts and problems, taking advantage of artificial intelligence and interpretability to improve decision making and obtain more efficient and effective results.

- **Constraints:** at the same time, this use case highlighted the constraints and difficulties in the use of post-hoc interpretability techniques. It is important to recognize that interpretability methods are not infallible and may sometimes provide approximate or partial results. Therefore, it is essential to take a critical and rigorous approach when interpreting and validating the outcome of interpretability techniques.

- **Combining AI models and interpretability modules:** this use case shows how the integration of AI models and interpretability modules can improve the predictive capability and understanding of models. This facilitates the adoption of AI-based solutions in business decision making.

- **Continuity in interpretability analysis:** finally, it should be emphasized that interpretability analysis should not be an isolated exercise applied during model development, but should be performed in a continuous, reproducible and reliable manner throughout the life of the model.

In conclusion, this artificial intelligence interpretability use case provided valuable experience in the implementation of AI and XAI techniques in a business context, and shows the potential of AI and interpretability to improve decision making, while revealing the limitations and difficulties associated with these techniques and the need for a critical and rigorous approach when interpreting and validating AI results.

Conclusion

"With the right programming, a computer can become a theater,
a musical instrument, a reference book, a chess opponent.
No other entity in the world except a human being has such an adaptable,
universal nature".

Daniel Hillis[74]

This study has presented Explainable Artificial Intelligence (XAI), its fundamentals, context and techniques for improving model interpretability. The main challenges facing artificial intelligence models in terms of interpretability and how technology can help to address them have been discussed, and a use case developed with ModelCraft™ has been shared to demonstrate how these techniques can be employed to understand and explain AI models.

The AI discipline, and within it XAI, has grown in importance worldwide in recent years as developing high-performance AI technologies has become a priority for many sectors, from health to security, financial services to energy and many others. Interpretability arises as the need to understand and improve AI models, which is particularly complex in the case of certain techniques.

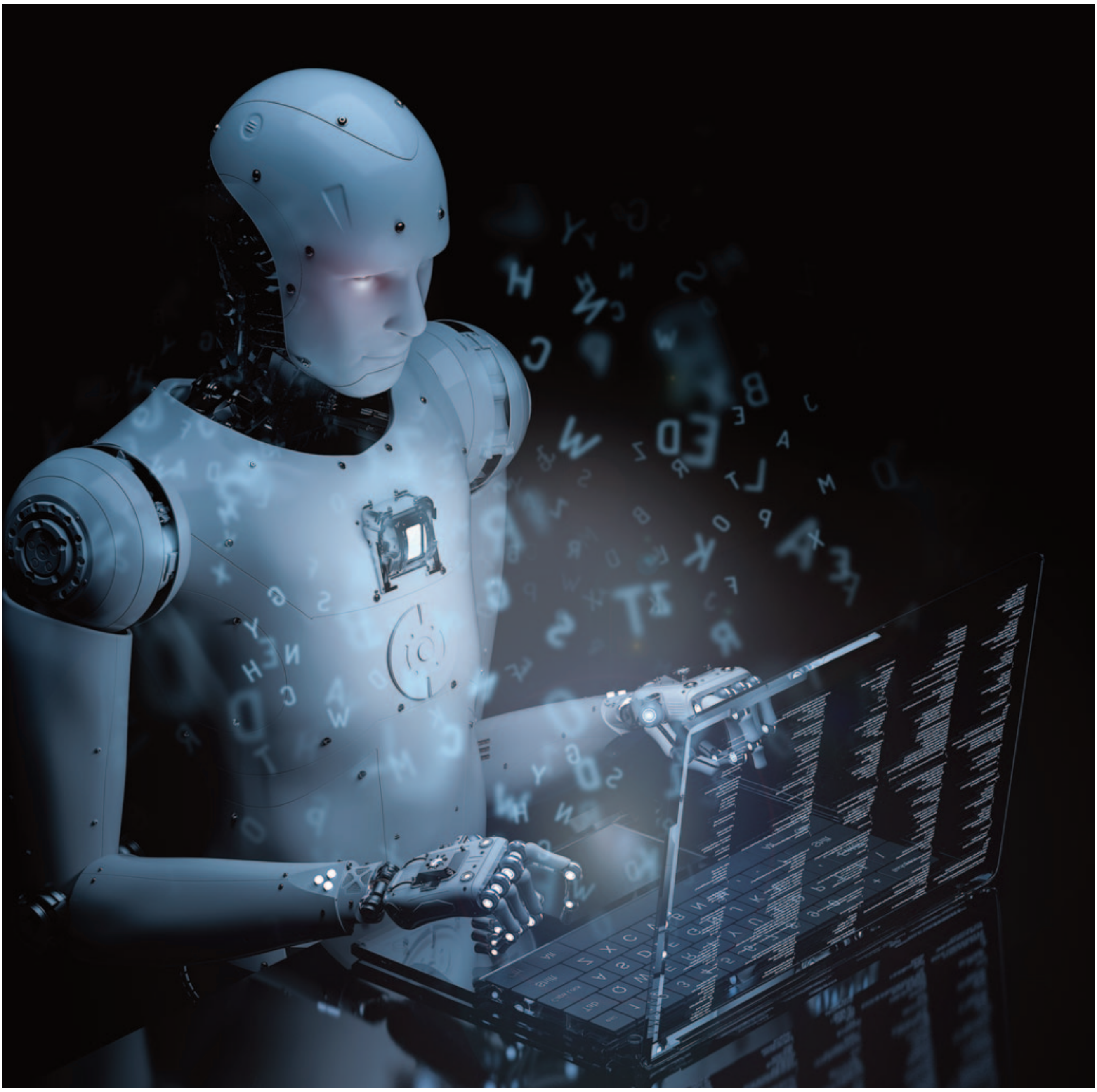
As seen, it can be difficult for AI models to explain their outcome or the logic behind their decisions. This is because these models use deep learning techniques and complex algorithms to learn from data, which are often difficult to interpret, and this poses challenges in evaluating AI models and the reliability of their output.

As a result, the AI regulatory framework is evolving rapidly, and organizations are expected to adapt to new requirements for transparency, explainability and fairness in the use of AI models. This implies the need for a comprehensive approach to integrate interpretability and explainability into each organization and its processes, encompassing interpretability techniques, model risk management, interdisciplinary collaboration and XAI training for professionals involved in AI development and implementation, among other areas.

In conclusion, the interpretability of artificial intelligence models is an emerging area of research, and it is expected to continue to develop and grow in importance as AI models become more complex, regulation continues to proliferate, and their use extends to more highly sensitive domains.

[74] Daniel Hillis (b. 1956), american inventor, entrepreneur and scientist, pioneer of parallel computing and its use in the field of artificial intelligence, with more than 300 published patents.

Glossary



Machine learning: **s**ubfield of artificial intelligence focusing on the development of algorithms and models that enable machines to learn and improve their performance on specific tasks through experience.

White box: AI system or model whose inner workings are simple to understand and explain.

Black box: an AI system or model whose inner workings are unknown or difficult to understand.

Right to an explanation: legal concept holding that individuals have the right to know how automated decisions affecting them are made and to receive an understandable explanation of how the algorithms involved work.

Explainability: AI system´s ability to provide clear and understandable reasons for its predictions or decisions to users and stakeholders. This involves providing detailed and contextualized information on how and why an AI model arrives at a particular conclusion, which promotes trust and makes it easier for the technology to be adopted.

GPT-4: fourth generation of the Generative Pre-trained Transformer model, developed by the OpenAI Foundation, which is used for natural language processing and text generation tasks.

Artificial intelligence (AI): field of study that seeks to develop systems capable of performing tasks that normally require human intelligence, such as learning, reasoning, perception and decision making.

Explainable artificial intelligence (XAI): AI approach that seeks to make artificial intelligence models more understandable and transparent to humans.

Interpretability: ease with which humans can understand an AI model's decision-making process, as well as the relationships between input features and predictions or decisions. An interpretable model allows users to discern how a specific prediction or decision is arrived at.

LIME (Local Interpretable Model-agnostic Explanations): an explainability technique that helps to understand the individual predictions of an AI model by creating local interpretable approximations.

Surrogate model: interpretable model that is trained to mimic the predictions of a complex and less interpretable AI model, such as a deep neural network. The goal of a surrogate model is to provide a simplified and understandable explanation of how the original model makes decisions.

Open AI Foundation: an artificial intelligence research and development organization, currently owned by Microsoft, whose stated goal is to ensure that AI benefits all of humanity.

Partial Dependence Plot (PDP): a visualization technique that shows the average effect of a feature on the predictions of an AI model, holding all other features constant. It helps to understand the relationship between features and predictions, and to detect potential interactions and nonlinearities.

Winograd Schema Test: natural language understanding test that assesses an AI's ability to resolve ambiguities in language through the use of common knowledge and reasoning.

General Data Protection Regulation (GDPR): European Union legislation that lays down rules for the collection, storage and processing of personal data of EU citizens.

AI bias: systematic bias present in training data or in the design of an AI algorithm that can lead to unfair or discriminatory decisions.

SHAP (SHapley Additive exPlanations): explainability technique that uses Shapley values from cooperative game theory to attribute the importance of each variable in an AI model's prediction.

Sparsity: model property whereby the model only considers the subset of variables that are really relevant for calculation.

Turing test: test proposed by Alan Turing in 1950 that evaluates a machine's ability to imitate human intelligence to the point of being indistinguishable from a human in a conversation.

Transformer: neural network architecture introduced by Google Brain in 2017 that is primarily used in natural language processing (NLP) tasks. Transformers are known for their ability to handle long data sequences and for their training efficiency. They are based on attentional mechanisms, which allow the network to weigh the relative importance of words or items in a sequence over time. Transformers have driven the development of state-of-the-art language models, such as GPT and BERT, and have revolutionized the NLP field.

Transparency: an AI system's openness and accessibility in terms of its design, structure, and internal processes. A transparent system allows users and stakeholders to examine and understand its components, algorithms and decisions.

Deep neural network: machine learning algorithm that has multiple layers of artificial neurons and is capable of learning hierarchical representations of data.

References

Broniatowski, D. (2021). Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. https://www.nist.gov/publications/psychological-foundations-explainability-and-interpretability-artificial-intelligence

Comisión Europea (2021). Artificial Intelligence Act / Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión. https://artificialintelligenceact.eu/

Comisión Europea (2019). Dirección General de Redes de Comunicación, Contenido y Tecnologías, Directrices éticas para una IA fiable, Oficina de Publicaciones, 2019, https://data.europa.eu/doi/10.2759/14078

C. Rudin, C. Chen, Zhi Chen, H. Huang, L. Semenova, C. Zhong. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. http://essay.utwente.nl/91965/

Doshi-Velez, F., et al. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608

Devis (2011). https://cs.nyu.edu/~davise/papers/WinogradSchemas/WSCollection.html

Dimensions (2022). https://app.dimensions.ai/discover/publication

EBA (2021). Discussion paper on machine learning for IRB models. https://www.eba.europa.eu/regulation-and-policy/model-validation/discussion-paper-machine-learning-irb-models

European Parliamentary Research Service (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)641530

Floridi et al. (2022). capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

Friedman, J. H. (2001). “Greedy function approximation: A gradient boosting machine.” Annals of statistics (2001): 1189-1232. https://www.jstor.org/stable/2699986

Gall, R. (2018). Machine Learning explainability vs interpretability: two concepts that could restore trust in AI, KDnuggets. https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

GDPR (2018), Recital 71. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Goldstein, A.; Kapelner, A.; Bleich, J; Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. https://arxiv.org/abs/1309.6392

Harnad, D. (2003). Can a machine be conscious? How? https://web-archive.southampton.ac.uk/cogprints.org/5330/

IBM (2022). Explainable AI (XAI). https://www.ibm.com/watson/explainable-ai

iDanae (2022). ML Applied to Credit Risk: building explainable models. Quarterly Newsletter 3Q22. iDanae Chair. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2022/10/Idanae-3Q22.pdf

Jonathon Phillips, P.; Hahn, H.; Fontana, P; Yates, A.; Greene, K. K.; Broniatowski, D. A.; Przybocki, M. A. (2021). Four Principles of Explainable Artificial Intelligence. NIST. https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence

Kaggle (2017). IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

LeCun, Y.; Bengio, Y.; Hinton, G. (2015). Deep learning. Nature. https://pubmed.ncbi.nlm.nih.gov/26017442/

Leventi-Peetz, A.-M., et al. (2022). Deep Learning Reproducibility and Explainable AI (XAI). https://arxiv.org/abs/2202.11452

Levesque, H. (2014). On our best behaviour. Written version of the Research Excellence Lecture presented in Beijing at the IJCAI-13 conference. Artificial Intelligence, vol. 212, pages 27-35. https://doi.org/10.1016/j.artint.2014.03.007

Lundberg, S. M.; Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. https://dl.acm.org/doi/10.5555/3295222.3295230

Management Solutions (2023). ModelCraft. Modelización por componentes. https://www.managementsolutions.com/es/microsites/soluciones-propietarias/modelcraft

Management Solutions (2022). Gamma. Sistema de gobierno de modelos. https://www.managementsolutions.com/es/microsites/soluciones-propietarias/gamma

Management Solutions (2021). Nota técnica sobre el EBA Discussion paper on machine learning for IRB models. https://www.managementsolutions.com/es/publicaciones-y-eventos/apuntes-normativos/notas-tecnicas-normativas/documento-de-debate-sobre-machine-learning-en-el-enfoque-irb

Management Solutions (2020). Auto machine learning, towards model automation. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/auto-machine-learning-towards-model-automation

Management Solutions (2018). Machine learning, a key component in business model transformation. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/machine-learning-a-key-component-in-business-model-transformation

Management Solutions (2015). Data science and the transformation of the financial industry. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/data-science

Marcinkevics, R. (2020). Interpretability and Explainability: A Machine Learning Zoo Mini-tour. ETH Zürich, Department of Computer Science, Institute for Machine Learning. https://arxiv.org/abs/2012.01805

McCarthy, J. (2004). What is artificial intelligence? Stanford University. http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell.2019,267, 1–38. https://www.sciencedirect.com/science/article/pii/S0004370218305988

OECD (2019). Principles for Artificial Intelligence. https://www.oecd.org/digital/artificial-intelligence/

Oneto, L., Chiappa, S., (2020). Fairness in Machine Learning. 2012.15816.pdf (arxiv.org)

Ribeiro, M. T.; Singh, S.; Guestrin, C. (2016). “Why Should I Trust You?” Explaining the Predictions of Any Classifier. https://arxiv.org/abs/1602.04938

Ribeiro, M. T.; Singh, S.; Guestrin, C. (2018). “Anchors: high-precision model-agnostic explanations”. AAAI Conference on Artificial Intelligence (AAAI).
https://ojs.aaai.org/index.php/AAAI/article/view/11491

Roscher, R.; Bohn, B.; Duarte, M.; Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries.
https://ieeexplore.ieee.org/document/9007737

Shapley, L. (1953). A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 307-317.
https://doi.org/10.1515/9781400881970-018

Sudjianto, A.; Knauth, W.; Singh, R.; Yang, Z.; Zhang, A. (2011). Unwrapping The Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification. Cornell University. https://arxiv.org/abs/2011.04041

Sudjianto, A.; Zhang, A. (2021). Designing Inherently Interpretable Machine Learning Models.
https://arxiv.org/abs/2111.01743

Turing, A. (1950). Computing Machinery and Intelligence. Mind 49: 433-460.

Vilone G., Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, vol. 76: 89-106.
https://www.sciencedirect.com/science/article/pii/S1566253521001093

White House OSTP (2022). Blueprint for an AI Bill of Rights.
https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Yang, Z.; Zhang, A.; Sudjianto, A. (2019). Enhancing Explainability of Neural Networks through Architecture Constraints. https://arxiv.org/abs/1901.03838

Zhou, N.; Zhang, Z.; Nair, V. N.; Singhal, H.; Chen, J.; Sudjianto; A. (2021). Bias, Fairness, and Accountability with AI and ML Algorithms. https://arxiv.org/abs/2105.06558

Note: This publication was produced with the help of several artificial intelligence (AI) tools. These tools were used for various tasks, such as searching for information, collecting and organizing data, and generating summaries. In any case, the final material in this publication was written by a person and not by an AI.

ManagementSolutions
Making things happen

Our aim is to exceed our clients' expectations, and become their trusted partners

Management Solutions is an international consulting services company focused on consulting for business, risks, organization and processes, in both their functional components and in the implementation of their related technologies.

With its multi-disciplinary team (functional, mathematicians, technicians, etc.) of more than 3,300 professionals, Management Solutions operates through its 44 offices (19 in Europe, 21 in the Americas, 2 in Asia, 1 in Africa and 1 Oceania).

To cover its clients' needs, Management Solutions has structured its practices by sectors (Financial Institutions, Energy, Telecommunications and other industries) and by lines of activity, covering a broad range of skills -Strategy, Sales and Marketing Management, Risk Management and Control, Management and Financial Information, Transformation: Organization and Processes, and New Technologies.

The R&D department provides advisory services to Management Solutions' professionals and their clients in quantitative aspects that are necessary to undertake projects with rigor and excellence through the implementation of best practices and the continuous monitoring of the latest trends in data science, machine learning, modeling and big data.

Javier Calvo Martín
Partner at Management Solutions
javier.calvo.martin@managementsolutions.com

Manuel Ángel Guzmán Caba
Partner at Management Solutions
manuel.guzman@managementsolutions.com

Segismundo Jiménez Láinez
Manager at Management Solutions
segismundo.jimenez@msspain.com

Luz Ferrero Peña
Supervisor at Management Solutions
luz.ferrero@msgermany.com.de

Management Solutions, Professional Consulting Services

Management Solutions is an international consulting firm whose core mission is to deliver business, risk, financial, organization, technology and process-related advisory services.

For further information please visit **www.managementsolutions.com**

Follow us at:

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña Málaga London Frankfurt Düsseldorf Paris Amsterdam Copenhagen Oslo Warszawa Wroclaw Zürich Milano Roma Bologna Lisboa Beijing Istanbul Johannesburgo Sydney Toronto New York New Jersey Boston Pittsburgh Atlanta Birmingham Houston San Juan de Puerto Rico San José Ciudad de México Monterrey Querétaro Medellín Bogotá Quito São Paulo Río de Janeiro Lima Santiago de Chile Buenos Aires