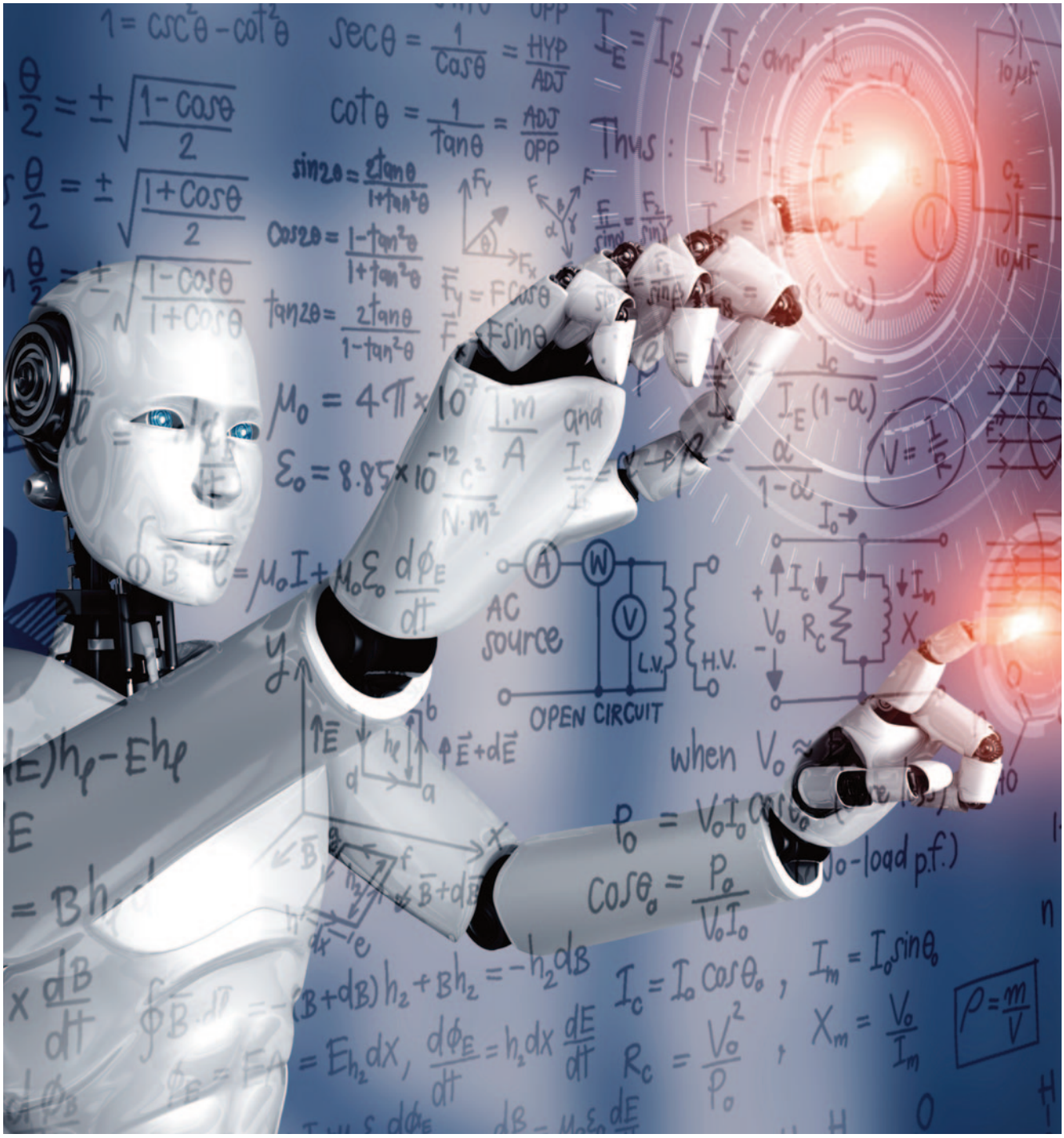
Executive summary

"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 [13].

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).