PRECISE - How to move from explainability to explanation: proposal for a process for explaining AI predictions

Vincent Lemaire, Marc-Eric Bobillier Chaumon, Suzie Grondin, Pierre Nodet, Thomas George, Nathalie Charbonniaud, Françoise Fessant, Moustapha Zouinar

PRECISE - How to move from explainability to explanation: proposal for a process for explaining AI predictions

Vincent Lemaire[1], Marc-Eric Bobillier Chaumon[2], Suzie Grondin[1,3], Pierre Nodet[1], Thomas George[1], Nathalie Charbonniaud[1], Françoise Fessant[1], Moustapha Zouinar[1]

[1] Orange Research, France
[2] Conservatoire National des Arts et Métiers, France
[3] Université Paris Sorbonne, France

Abstract. In the case of decisions made by AI, this article explores how to shift from a technology-centric approach to a human-centric approach to explainability, providing understandable and contextualized explanations to end users, particularly in the workplace. The article highlights the importance of developing explainability mechanisms tailored to user profiles, needs, and contexts, adopting a collaborative, dynamic, and contextualized approach. It proposes a structured process, inspired by the CRISP-DM model, comprising the following steps: context analysis, needs analysis/gathering, modeling, implementation, integration, and validation. This process, called "PRECISE", aims to ensure that explanations are relevant, understandable, and actionable, while avoiding bias and building trust. The approach encourages close collaboration between technical stakeholders and professional users. An illustrative example is also given in the case of fraud detection.

1 Introduction

Machine learning, one of the branches of artificial intelligence (AI), has enjoyed considerable success in recent years. The decisions made by these models are increasingly accurate, but also increasingly complex. Some of these models are akin to black boxes: their decisions are difficult, if not impossible, to explain [7] in a way that is understandable to users. This lack of explainability can lead to several undesirable consequences: mistrust, lack of confidence or overconfidence on the part of the user, reduced usability of the models, presence of bias, etc. It is from these needs that the field of eXplainable AI (XAI) was born. XAI [1, 40] aims to develop methods and tools to make the decisions made by machine learning models understandable to users.

Despite progress, the "black box" nature of many artificial intelligence systems remains a significant barrier to adoption and acceptance, particularly in sectors such as medicine, finance, law, and fraud detection, where decision-making must be transparent and verifiable. These systems, particularly those based on

deep learning[4], often function as decision-making processes whose logic is beyond the understanding of users [11,39]. This opacity limits users' ability to interpret and understand the constructed knowledge, and to trust or exercise control over the decisions made by AI — particularly concerning the individual's own value system.

A lack of clear, accessible explanations can limit trust, which is essential for integrating AI into professional activities [20]. Without understandable explanations, users may mistrust or reject these systems [25]. The difficulty of understanding how models work can also mask the reproduction of biases [37], which complicates ethical management and allocating responsibility in the event of failure [16]. Therefore, explainability systems are crucial for interpreting AI predictions for several fundamental reasons (non-exhaustive list). These reasons can be organised into two categories: "user trust" (1–6) and "model uncertainties/limitations/bias" (7–8):

* Building trust [4]: Explainability makes AI decision-making processes understandable, which promotes user trust. Without clear explanations, users may sometimes doubt or mistrust the automatic decisions made by AI when they have access to them[5].
* Ensuring transparency and accountability [14]: In critical environments such as nuclear power stations, the military, cybersecurity and medicine, it is crucial to understand how and why decisions are made, particularly in order **to justify or control** them.
* Facilitating adoption and acceptance [6,34]: Tailoring explanations to users' needs and co-designing them with users enables better adoption of the tool, avoiding mistrust or suspicion. Adoption **improves** users' decision-making abilities. They also make it possible to quantify the intuitions of business teams.
* Supporting the co-construction of meaning: Explanations are not merely technical; they must be contextualized and cooperative, enabling users to understand, **control**, and adjust AI in their activities. They must also enable the potential **discovery** of new knowledge [30].
* Facilitate **compliance** with regulations [18] and ethical requirements regarding the use of AI.
* Ensure greater **accountability** by enabling the identification of the causes of automated decisions. AI systems should be subject to explanation standards similar to those currently applied to humans [18]. AI provides support at times, but it remains on the sidelines at other times and must not erase human autonomy [8,46]. There is also the idea that the system and explanations must not only be flexible (adapted to the situation) but also adjustable. In other words, they should be able to learn from their mistakes, from the situation, and from the user to evolve and improve...

[4] Often considered "black box" methods

[5] In some use cases (Waze, for example), users sometimes do not have access to the "decisions" made by AI and therefore do not have the opportunity to question them.

* Managing uncertainty and limitations: Explainability is vital in highlighting the limitations of the system, particularly in cases of uncertainty or lack of data, to avoid erroneous or risky decisions [52].
* Improve the **reliability** of models by enabling more effective detection of biases [48] or errors.

Explainability is necessary for AI to be integrated in an ethical, reliable, and acceptable manner in contexts where safety, accountability, and trust are paramount [5]. Consequently, this article suggests that for AI to be truly integrated and used reliably, contextualized, cooperative, and adaptive explainability mechanisms must be developed [4] to enable users to understand and interact effectively with these systems [49]. To this end, and to ensure that explainability is not solely the result of a "push" of technology by the "data scientists", we propose a "pull" type "process" that will hopefully enable us to move from explainability to explanation (or co-explanation, to emphasize the importance of the collective), which is rooted in the reality of work activities and practices

Note: in this article, our focus is primarily on the predictive case, where an AI model must predict a class or a numerical value. We therefore have the model's prediction and its confidence score, as well as the actual class or value that it should have predicted. However, we believe that the points we raise can be generalised to other situations.

2 The different forms of explainability

Forms of explainability differ according to their level of detail (summary or in-depth), their format (textual, visual, interactive), their suitability for expert users (in data science or the field of application) or not, and their purpose (to understand, justify, control, act, explore). In this article, we advocate a contextualized, cooperative, and dynamic approach that considers the activity, context, and specific needs of users in order to produce explanations that are truly useful and operational in professional environments.

The state of the art offers several forms and dimensions of AI explainability. While our approach differs from the technical taxonomy proposed in [44], it can be said that these forms vary depending on the context, audience and objectives. In particular, we can cite the following (non-exhaustive list):

1. Technical explainability (transparency) [19] to enable experts or designers to understand how the system arrives at a decision.
2. Explanation for understanding (explainability for the user) [15, 53] to promote trust, control, and ownership by the end user.
3. Accountability explanation [17]: to ensure responsibility and transparency in decision-making, it is important to establish who (human, machine or program) is responsible for what.
4. Summary explanation vs. detailed explanation [31] depending on the desired level of analysis, global explanation at the model level or local explanation for a specific prediction.

5. Contextual explanation [3] tailored to the user's profile with the aim of making the explanation relevant, understandable, and useful to the recipient.
6. Interactive explanation and dialogue [26] to promote active understanding and more effective learning.
7. Counterfactual explanation [21] to provide answers such as "What would need to be changed to achieve a different result?"
8. Visual explanation [22,23] and simplified explanation that will be useful for non-experts or in contexts where speed of understanding is crucial.
9. Actionable explanation [43] that provides concrete recommendations or actions to take based on the explanation.

In order to achieve the goal of a contextualised, cooperative and dynamic approach, it is clear that several of the above forms may need to be combined. We must also define what we mean by "user" of the explainability form produced. The end user is not necessarily unique, and can sometimes obscure the bigger picture. They may only be the first link in a chain of decisions that must be taken into account, as in the case of Électricité de France (EDF), as described in [4], where there is a diversity of profiles involved.

3 Proposal for the "PRECISE" process

In order to achieve the objective of a contextualised, cooperative and dynamic approach, the process to be implemented may resemble that of the CRISP-DM process [13]. This approach combines user-centred design with the CRISP-DM process. The aim is to guard against automation biases in terms of explainability, particularly to ensure that users do not place undue trust in artificial intelligence algorithms. The process will seek to avoid providing explanations that are inadequate for users' needs in the use case concerned. It is important not to create an excessive workload through additional checks, or to generate confusion or loss of confidence. This requires clearly identifying the actual conditions under which the work activity is carried out, such as the task sequence, individual and collective decision-making procedures, contingencies, and circumstances of situations. It also requires adapting to the resources and cognitive logic of individuals who will handle the provided explanatory elements. These elements should be directly usable without any additional steps required to understand them; if not, we can assume that the process has failed. It is only through the full implementation of this process that we can move from explainability (often techno-centred) to explanation (human-centred).

Indeed, the initial techno-centric approach often focuses on model transparency (black box $\rightarrow$ technical explainability). The aim is to make the system understandable to expert data scientists. However, it is not well suited to end users as it provides little context. Therefore, it is up to end users to evaluate the reliability of the explanations provided and adapt them to their own activities if necessary.

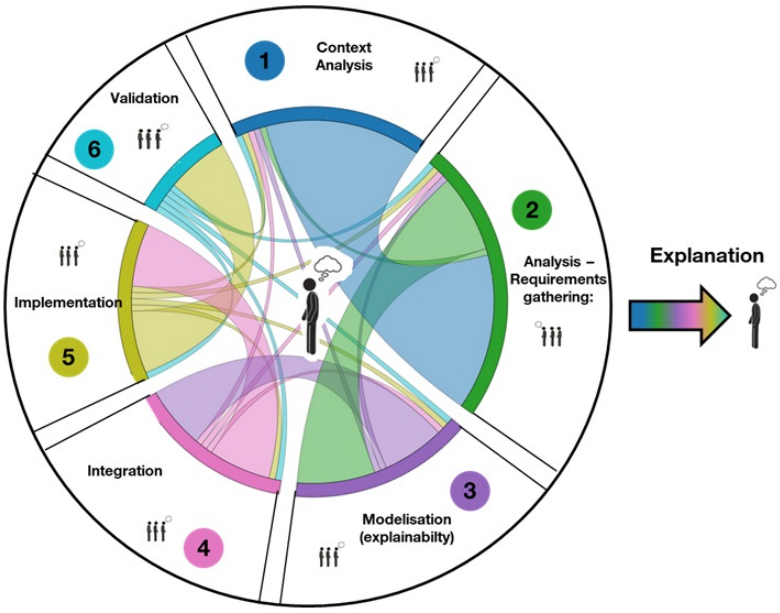

Fig. 1. Chord diagram process: from explainability (technology-centred) to explanation (human-centred). At each of the stages mentioned here, the end user(s) and technical stakeholders consult each other to determine the appropriate method for meeting the needs of the given context. It is possible to move from one stage to the next, or to return to a previous stage, at any time if necessary. However, it is not advisable to skip a stage (for example, there is no connection shown in the diagram between stages 2 and 4). The time required for each step is not necessarily the same. Here, for illustrative purposes, the time taken by the first two steps has been emphasised. **All six steps** lead to the **explanation**, which is the **result** of the entire process.

Consequently, the end user(s) must be involved in creating the explanation from start to finish. They should be involved from the outset to provide the specifications, at certain key stages, and finally as the final recipient for validation. A 'go/no-go' stage must then be agreed upon for acceptance of the delivery ('go') or a 'no-go', in which case the process returns to a previous stage or even the beginning. As the user's knowledge has evolved, they may now be better able to describe their needs.

Like the CRISP-DM process [13, 28], the proposed process (see Figure 1) will involve several professions (non-exhaustive list): end users (who may have diverse fields of knowledge); a data scientist (who is familiar with techno-centric explainability methods); an ergonomist (who reports on conditions, methods and variability of situations); a member of the information system (who indicates availability of data and computing resources); and any other relevant actors.

Note: The process proposed in this article is part of a logic of integration with the CRISP-DM model, particularly during the latter's modeling and explanation

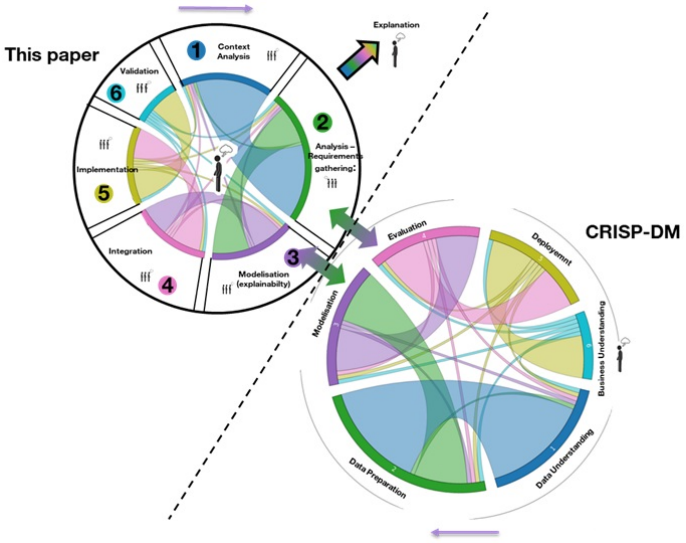

Fig. 2. Entanglement of chord diagrams from the PRECISE and CRISP-DM processes.

phases (see Figure 2), like two wheels turning in synergy: during the CRISP-DM modeling phase, the choice and adjustment of explainability models are carried out. This then feeds directly into the PRECISE modeling phase, which aims to make these models understandable and contextualized for end users. This articulation ensures that the explainability methods are not only technically appropriate, but also aligned with the real needs of CRISP-DM model users, taking into account the context of use and varied profiles. The cooperative and dynamic approach proposed in the document thus promotes a continuous loop between these two stages, reinforcing the relevance and effectiveness of the explanations. One consequence of this interconnection is feedback from the PRECISE model to the CRISP-DM model: users sometimes have to make compromises, such as choosing a classification model (from the CRISP-DM model), in order to achieve their objectives within the PRECISE model.

3.1 Context analysis: The first step in the process is to consider the context in which AI predictions are explained. This involves first understanding who needs explanations, in what context, and for what activities. Next, it is necessary to analyze the organizational, technical, and social constraints that shape this use, while anticipating risks, trust and control needs, and adapting the form and scope of the explanation to these parameters. This approach enables explanations to be designed that are relevant and consistent with the activity in question and its challenges. A non-exhaustive list of questions to consider when thinking about the context of use could include: Who are the stakeholders and what are their profiles? What is the organisation's culture? What are their specific explanation needs? What is the nature of the activity? What organisational, regulatory and social issues are at stake? How does the activity unfold over time

and space? What physical, technical and social factors are involved? What is the working environment (control room, field, cockpit, nuclear power plant, etc.)? Which contextual elements (goals, interaction with the environment, stress, fatigue, etc.) could influence the perception and interpretation of explanations? How will the activity evolve? How should the explanation fit into this dynamic?

3.2 Analysis - Requirements gathering: This is the stage at which explanation requirements are defined. When designing explanations for the predictions of an AI system, it is important not to focus solely on one user or recipient. In a given context, several individuals may require an explanation, each with different needs, expectations and levels of understanding. For instance, in an industrial or medical setting, explanations may be intended for field operators, supervisors, engineers, regulatory officials, or strategic decision-makers. These groups have different cognitive profiles, technical skills and challenges, which influence the nature, form and depth of the required explanations. Therefore, it is crucial to adopt a pluralistic approach, identifying all the actors potentially involved in the decision-making or operational process and adapting the explanations to their profiles and the context in which they will be used. This ensures that each recipient receives information that is relevant, understandable and useful, thereby promoting trust in, and ownership of, the AI system, and enhancing its effectiveness in real-world applications. Explanations must be designed as part of a systemic approach that takes into account the diversity of actors and the complexity of situations, rather than focusing solely on the user. These defined needs will then guide the design of the necessary explanations.

3.3 Modeling: Should we choose an off-the-shelf explainability method or design one [2]? We are currently in the technical (technocentric) phase. The aim is to find an off-the-shelf explainability method that meets the identified needs. If none exist, a 'research' phase may be necessary to develop one. When we say 'a' method, we may in fact need to consider several methods to fully cover the user's needs, either by combining them or adding them together. A classic example is variable importance, where we can combine global and local methods (e.g. for decision trees, see [31]). Modeling may also require additional understanding or knowledge discovery. For instance, counterfactuals may be sought not to explain the model's decision, but rather to challenge it or verify post hoc whether the labels in a classification problem can be questioned. For example, is there a false positive or was the label incorrect?

3.4 Implementation: This step involves the technical implementation of the selected explainability method. This involves coding, developing, configuring and practically implementing the method with the aim of creating an operational version that can function within the user's existing system. Ideally, all the necessary elements for implementation will already be in place. Otherwise, it will be necessary to add hardware and/or software, as well as a connection to external sources of useful information. It is important to consider what end users deem possible or impossible to add or include here. During this stage, it may become apparent that the AI model is not immutable. It may even be called into ques-

tion if it is found that it will not meet the expressed needs due to the available explainability methods and the computing time that would be required. In this case, the decision may be made to reduce decision-making capacity in order to improve interpretability.

3.5 Integration: Integration refers to the process of incorporating this method or module into an operational environment. The aim is to ensure that the explainability method works in harmony with other components, processes or interfaces. This involves ensuring that the explainability method fits effectively into the workflow and existing architecture, and that it is accessible to and usable by end users and other modules within the system. Technical integration is accompanied by an assessment of the conditions for acceptance of these explanations in the real-world context of the activity. In other words, we assess the extent to which the explanations provided improve, degrade or reconfigure users' activities. This is because adoption does not depend on individuals' prior favourable perceptions of AI and its explanations (following training or awareness-raising), but rather on this system's actual favourable effects on users' practices and systems of activity (social, organisational and professional).

NB: The evaluation of explainability methods [33] is an open research question [36], with evaluation criteria including, for example, faithfulness (i.e. the extent to which the explanation is representative of the AI system), readability, and plausibility (i.e. the extent to which the explanation convinces the user). This point could potentially be an additional step in the process proposed in this article, though this is not developed here.

Reasons for failure or success in steps 4 and 5[6]: Although specifications should have been established by the end of the first two stages to ensure the project's success, there are still potential failure points in the implementation and integration phases.To help anticipate the most common reasons for failure, we will cite the following examples, which should therefore have been considered at the start of the project:

- (i) Technical incompatibility: Difficulty integrating the explainability method into the existing architecture (software compatibility, compatibility with other modules).
- (ii) Resource limitations: Lack of time, technical skills, or financial resources to carry out the implementation.
- (iii) Algorithmic complexity: The explainability method is too computationally or memory-intensive, making its deployment impractical in real time or on limited systems.
- (iv) Insufficient or inadequate data: The method requires specific data or metadata that is unavailable or of poor quality, preventing reliable explanation.
- (v) Opacity or technical difficulty: The explainability method is difficult to code or operate in practice, particularly if it relies on experimental or unstable techniques.

[6] Steps 3, 4, and 5 are quite intertwined, so there may be work cycles between these three steps and the consolidation of results.

- (vi) Security or confidentiality risks: The method could reveal sensitive information or compromise system security, which would prevent its implementation.
- (vii) Organizational or regulatory tensions: Regulatory constraints or internal resistance could prevent the method from being integrated, particularly if it does not comply with standards or policies.
- (viii) Interoperability issues: Difficulty in communicating the new explainability method with other existing components or systems (databases, interfaces, monitoring modules), leading to disruptions or inconsistencies in the flow of information.
- (ix) Lack of complete mastery (or overconfidence) for some of the building blocks: complete mastery (or overconfidence) cannot always be guaranteed for all components, as the objective is to deploy the system as a whole, including modules from partners or pre-existing solutions.
- (x) Non-acceptance of the system or explanations in the context of the activity being carried out: excessive workload related to the analysis and evaluation of explanations; proposals/suggestions that contravene business rules and professional quality criteria; individual evaluation of explanations when the activity is carried outcooperatively, etc

3.6 Validation and evaluation tests: The aim of this step is to ensure that the method works as intended in the target context. This is achieved by verifying the quality, relevance and reliability [41, 51] of the explanations provided. The validation step differs from the implementation and integration steps in the following ways, for example. Field tests are carried out to verify whether explanations are understandable, useful and tailored to user profiles (e.g. experts, non-experts and operators). These tests measure comprehension, confidence and satisfaction. The validation process is not limited to technical or functional verification, but also includes a qualitative and contextual assessment of the explanations' quality. The consistency between the model and its explanations is validated, often through auditing methods or comparison with expert knowledge. Verification that the explanations improve users' understanding, confidence and ability to act within their context. Validation is not limited to technical verification; it also assesses the effect of explanations on activity or performance. This potentially includes a regulatory compliance or ethical assessment step, which is not necessarily covered in technical implementation or integration. Additionally, there is an ethical compliance stage, as explanations must correspond to and integrate with an existing professional community with collectively predefined and accepted business rules, values, ways of doing things and quality criteria.

4 Illustration - Example of an expert "accompanied" by AI in the TRIO platform

4.1 The TRIO platform and the case of the "wholesale market"

Telecommunications companies in different countries use various international routes to exchange traffic. In a 'wholesale market', telecommunications operators can negotiate with other operators to obtain traffic to offset a deficit or send traffic to other routes. Minute exchanges enable operators to buy and sell terminations. Prices on the wholesale market may vary daily or weekly. Operators seek the lowest-cost routing function to optimise their exchanges on the wholesale market. However, the quality of routes on the wholesale market may also vary, as traffic may take an indirect route.

A value chain exists between the operators that connect two customers. However, this can be disrupted if a fraudster finds a way to communicate without paying. Over the past few decades, fraud has become an increasing concern within the telecommunications industry. A 2017 CFCA survey [12] estimates global annual losses due to fraud at $30 billion (USD). Therefore, it is essential to detect and prevent fraud in this area as far as possible. In the wholesale market, operators maintain a list of known frauds [24] and fraud detection platforms already exist.

One such platform was set up by Orange (operating as a wholesale provider) in 2019. Called TRIO (short for 'TRaffic Investigation at Orange'), this platform contains modules that use information provided by experts, such as the scoring module (a classifier), while others explore data to interact with experts by identifying new patterns, including malicious ones[7]. This objective is achieved through knowledge discovery and active learning techniques. These exploration modules are designed to adapt to the constantly changing behaviour of fraudsters, within the constraint of the limited time that experts can devote to this exploration (in the case of this platform). This platform is similar to the one presented by Veeramachaneni et al. in [47]. Both platforms combine a supervised model for predictions with unsupervised models for exploring unknown patterns and consider user feedback during the learning phase.

4.2 The need for explanations, the context, and the actors involved

In the case of fraud involving international calls[8], the platform consists of a machine learning model [10] that classifies examples as 'normal' or 'fraudulent'. This model is based on a particular type of AI called 'end-to-end frugal automl', as described in Boulle et al. (2025). This model enables the automatic analysis of the database of past calls. When fed with qualified examples, it calculates a probability rate or score. Initially, the number of qualified examples (labelled by

[7] One of the uses of this platform is described here: AI is a game changer are you ready?

[8] Detecting fraud on international calls with AI

an expert) was relatively small. The data analyst focuses on the model's results with the highest scores and confirms whether they are fraudulent, enabling the blocking of these phone numbers and the prevention of further fraud. From the more than 200 million international call reports analysed every day, the machine learning model generates approximately 65,000 alerts.

In general, fraud experts are responsible for developing tools to enable fraud detection. In this case, based on interviews conducted at the start of the project, the fraud expert requested a list of features to validate and supervise the machine learning model(s) (the detection tool(s)).

The first feature was a human-machine interface (HMI) to view call groups that had been detected as either fraudulent or non-fraudulent, in order to confirm or overturn the AI model's decision. While not strictly an explainability module, it was nevertheless an essential prerequisite for the expert, particularly in the initial phase of the detection platform's development. The HMI also enabled the continued collection of labels 'qualified' by the expert, thereby increasing the size and diversity of the labelled examples. As the basic learning set grew over time, the machine learning model could be updated regularly (a feature that had been requested at the start of the project).

The explanatory needs of data analysts, who are also experts in this type of fraud, have evolved over time and can be categorised into four types:

1. (2020) having confidence that examples scored with high scores are indeed fraudulent cases (with a high degree of confidence);
2. (2020) having "global" information about the importance of the variables present in the scoring model input;
3. (2025) having "local" information on the importance of the variables present in the scoring model input for each case taken individually;
4. (2026) having counterfactuals of examples predicted as fraudulent (or non-fraudulent);

The process illustrated in Figure 1 was therefore repeated three times in 2020, 2025, and early 2026. The stakeholders involved in designing all of these features were: ergonomists and developers for the HMI, fraud experts, data scientists for machine learning model training, and AI researchers (see below for the reason for the latter). The responses to the four aforementioned requests are described below.

4.3 Implemented or developed (technology-centered) explainability methods

Confidence in the scores produced - Implementation of a continuous monitoring process in which high-scoring cases are regularly re-evaluated and models are adjusted based on new data and feedback.

Overall importance of classifier input variables: The importance of the classifier's input variables for predicting fraud indicates the extent to which

each variable contributes to the classification model's performance. It identifies the variables with the greatest impact on the model's decision-making process, highlighting the most relevant factors for prediction. Understanding this helps to interpret the model and improve understanding of the factors influencing classification. In the use case described here, these values provide a 'synthetic' view of the most influential factors for a model training population. Initially (in 2020), the importance values provided to users were taken from [9]. This type of information has long been considered a 'minimum viable product' (MVP) of XAI.

Local importance of input variables to the classifier - Local importance focuses on a single, specific prediction, evaluating how each feature influences the model's decision for that particular example. This allows each variable's contribution to be analysed within the context of a specific observation, providing a more detailed and personalised explanation of the model's decisions. For instance, local importance could reveal that the most decisive feature in classifying a particular transaction as fraudulent is its unusual geographic location. Recent state-of-the-art research has presented two popular methods: LIME [38] and SHAP [42]. In the case of the industrial process described here, users initially opted for a Shapley-based calculation. However, it was initially impossible to implement this method due to the excessively long computation times observed with the KernelShap library [32]. The decision was therefore made to 'go back to research' and see if it was possible to produce an analytical method that respected the computation time constraints with the classifier integrated into the Khiops library. This was achieved and the results were published in 2023 [27].

Counterfactuals of non-fraudsters - Among XAI methods, counterfactual reasoning is a psychological and sociological concept [35]. It involves examining possible alternatives to past events [29, 45, 50]. Humans often use counterfactual reasoning to imagine what would have happened if an event had not occurred; this is precisely what counterfactual reasoning involves. When applied to artificial intelligence, questions might be asked such as, "Why did the model make this decision rather than another?" or "How would the decision have been different if a certain condition had been modified?" In the case of fraud, the aim was to explore instances where the classifier predicted 'no fraud' despite a very close counterfactual existing, thus generating new insights. The decision was made to conduct further research to see if it was possible to produce an analytical method for calculating counterfactuals that respected the computation time constraints in the case of the classifier integrated into the Khiops library. This was achieved and the results were published in 2024 [30].

4.4 Integration, implementation, validation and evaluation testing

Meticulous observation of the integration and implementation was conducted at each stage of the process to ensure it precisely met the initial expectations. All stakeholders collaborated closely with experts to verify that every technical and

functional detail aligned with their vision and requirements. Structured interviews were conducted to reinforce this validation, during which fraud experts could express their impressions, ask questions and confirm that the final result met their expectations. This iterative approach ensured optimal consistency between the design, the implementation and the established objectives, while also fostering transparent communication and enabling continuous adaptation as needed.

5 Conclusion

This article emphasises the importance of shifting from technical explainability to a human-centred approach to explanation in the context of artificial intelligence in the workplace. The article proposes a structured process inspired by the CRISP-DM model to emphasise the need to integrate end-user needs, context and profiles from the design stage onwards, ensuring explanations are relevant and effective. The proposed approach emphasises a cooperative, dynamic and contextualised methodology, enabling the avoidance of automation bias, the strengthening of trust and the ethical and responsible use of AI systems. The success of this transition ultimately hinges on close collaboration between the various stakeholders involved and rigorous validation of explanations within their intended context. Implementing this process is an essential step towards making artificial intelligence more transparent, understandable and accepted in professional environments.

References

1. Allen, G.I., Gan, L., Zheng, L.: Interpretable machine learning for discovery: Statistical challenges & opportunities. Arxiv preprint:2308.01475 (2023)
2. Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., et al.: One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012 (2019)
3. Avgerou, C.: Contextual explanation. MIS Quarterly **43**(3), 977–A18 (2019)
4. Bennani, R., Bobillier Chaumon, M.E., Fréjus, M.: Doctoriales arpege 2025: Concevoir une IA explicable, appropriable et de confiance: approche située par et pour les métiers. In: Doctoriales (2025)
5. Bennani, R., Fréjus, M., Bobillier Chaumon, M.E.: De l'explicabilité à l'explication des IA : perspective située incarnée par et pour les métiers. In: Conférence EGC - Atelier EXPLAIN'AI 2025. GDR Radia-CNRS and Association EGC, Strasbourg, France (Jan 2025), `https://hal.science/hal-04920628`
6. Bobillier Chaumon, M.E., Bennani, R., Mourgaud, L., Frejus, M., Lemaitre, D., Grisvard, O.: Des ia explicables pour des technologies acceptables et soutenables pour le travail : Enjeux et contours de l'explication dans et pour l'activité. La Revue des Conditions de Travail. Anact (2026)
7. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S.: Benchmarking and survey of explanation methods for black box models. Data Mining and Knowledge Discovery **37**(5), 1719–1778 (2023)

8. Bonneau, C.: Conceptualiser l'articulation technologie-organisation dans une perspective communicationnelle: entretien avec carole groleau. COMMposite **13**, 86–110 (01 2010)
9. Boullé, M.: Compression-based averaging of selective naive bayes classifiers. Journal of Machine Learning Research **8**(58), 1659–1685 (2007), http://jmlr.org/papers/v8/boulle07a.html
10. Boullé, M., Voisine, N., Guerraz, B., Hue, C., Olmos, F., Popescu, V., Gouache, S., Bouget, S., Bondu, A., Gauthier, L.A., Benrekia, Y.N., Clérot, F., Lemaire, V.: Khiops: An end-to-end, frugal automl and xai machine learning solution for large, multi-table databases (2025), https://arxiv.org/abs/2508.20519
11. Burrell, J.: How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society **3**(1), 2053951715622512 (2016)
12. CFCA: 2017 Global Fraud Loss Survey. Survey Results, Communications Fraud Control Association (2018)
13. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: Crisp-dm 1.0 step-by-step data mining guide. Tech. rep., The CRISP-DM consortium (August 2000), http://www.crisp-dm.org/CRISPWP-0800.pdf
14. Chapuis, V., Guégan, D.: Pour une conception et une utilisation responsables de l'Intelligence Artificielle. In: ERGO'IA 2023. Bidart, France (Oct 2023), https://hal.science/hal-04405065
15. Chromik, M.: Human-centric explanation facilities: Explainable AI for the pragmatic understanding of non-expert end users. Ph.D. thesis, Dissertation, München, Ludwig-Maximilians-Universität, 2021 (2021)
16. Cihon, P., Schuett, J., Baum, S.D.: Corporate governance of artificial intelligence in the public interest. Information **12**(7) (2021)
17. Diakopoulos, N.: Accountability, transparency. The Oxford handbook of ethics of AI **17**(4), 197 (2020)
18. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C.T., Gershman, S.J., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., Weller, A., Wood, A.: Accountability of ai under the law: The role of explanation. ArXiv **abs/1711.01134** (2017), https://api.semanticscholar.org/CorpusID:2092882
19. Ehsan, U., Liao, Q.V., Muller, M., Riedl, M.O., Weisz, J.D.: Expanding explainability: Towards social transparency in ai systems. In: Proceedings of the 2021 CHI conference on human factors in computing systems. pp. 1–19 (2021)
20. Gamkrelidze, T., Zouinar, M., Barcellini, F.: The "Old" Issues of the "New" Artificial Intelligence Systems in Professional Activities, chap. 6, pp. 71–86. John Wiley & Sons, Ltd (2021)
21. Guidotti, R.: Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery **38**(5), 2770–2824 (2024)
22. Halnaut, A.: Méthodes et outils d'analyse visuelle pour la compréhension, l'optimisation et l'élaboration de modèles de réseaux de neurones profonds. Ph.D. thesis, Université de Bordeaux (Mar 2024), https://theses.hal.science/tel-04633751
23. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: European conference on computer vision. pp. 3–19. Springer (2016)
24. I3 Forum: I3f Fraud Classification. White paper 3, I3Forum (May 2014)
25. Kizilcec, R.F.: How much information? effects of transparency on trust in an algorithmic interface. In: Proceedings of the 2016 CHI Conference on Human Factors in

Computing Systems. p. 2390–2395. CHI '16, Association for Computing Machinery (2016)
26. Lakkaraju, H., Slack, D., Chen, Y., Tan, C., Singh, S.: Rethinking explainability as a dialogue: A practitioner's perspective. arXiv preprint arXiv:2202.01875 (2022)
27. Lemaire, V., Clérot, F., Boullé, M.: An efficient shapley value computation for the naive bayes classifier. In: Meo, R., Silvestri, F. (eds.) Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2023, Turin, Italy, September 18-22, 2023, Revised Selected Papers, Part I. Communications in Computer and Information Science, vol. 2133, pp. 75–90. Springer (2023)
28. Lemaire, V., Clérot, F., Voisine, N., Hue, C., Fessant, F., Trinquart, R., Olmos Marchan, F.: The data mining process : a (not so) short introduction (2017), https://www.researchgate.netpublication/313528093_The_Data_Mining_Process_a_not_so_short_introduction
29. Lemaire, V., Hue, C., Bernier, O.: Data Mining in Public and Private Sectors: Organizational and Government Applications, chap. Correlation Analysis in Classifiers, pp. 204–218. IGI Global (2010)
30. Lemaire, V., Le Boudec, N., Guyomard, V., Fessant, F.: Viewing the process of generating counterfactuals as a source of knowledge: a new approach for explaining classifiers. In: 2024 International Joint Conference on Neural Networks (IJCNN) (2024)
31. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. Nature machine intelligence **2**(1), 56–67 (2020)
32. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
33. Martens, D., Shmueli, G., Evgeniou, T., Bauer, K., Janiesch, C., Feuerriegel, S., Gabel, S., Goethals, S., Greene, T., Klein, N., Kraus, M., Kühl, N., Perlich, C., Verbeke, W., Zharova, A., Zschech, P., Provost, F.: Beware of "explanations" of ai (2025), https://arxiv.org/abs/2504.06791
34. Mathew, D.E., Ebem, D.U., Ikegwu, A.C., Ukeoma, P.E., Dibiaezue, N.F.: Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of ai models for human. Neural Processing Letters **57**(1), 16 (2025)
35. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
36. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Comput. Surv. **55**(13s) (Jul 2023)
37. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, USA (2016)
38. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144 (2016)
39. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**, 206 – 215 (2018), https://api.semanticscholar.org/CorpusID:182656421

40. Saeed, W., Omlin, C.: Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems **263**, 110273 (2023)
41. Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., Vössing, M.: A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. p. 617–626. AIES '22, Association for Computing Machinery (2022)
42. Shapley, L.S., Shubik, M.: A method for evaluating the distribution of power in a committee system. American Political Science Review **48**(3), 787–792 (1954)
43. Singh, R., Miller, T., Lyons, H., Sonenberg, L., Velloso, E., Vetere, F., Howe, P., Dourish, P.: Directive explanations for actionable explainability in machine learning applications. ACM Transactions on Interactive Intelligent Systems **13**(4), 1–26 (2023)
44. Sokol, K., Flach, P.: Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 56–67. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372870, https://doi.org/10.1145/3351095.3372870
45. Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11781–11803 (2021). https://doi.org/10.1109/ACCESS.2021.3051315
46. Suchman, L.A.: Plans and situated actions: the problem of human-machine communication. Cambridge University Press, USA (1987)
47. Veeramachaneni, K., Arnaldo, I., Bassias, C., Li, K., Cuesta-Infante, A.: AI^2: Training a Big Data Machine to Defend. In: IEEE International Conference on Intelligent Data and Security (IDS). pp. 49–54 (Apr 2016)
48. Virtanen, K.: Using xai tools to detect harmful bias in ml models (2022)
49. Vuarin, L., Steyer, V.: Le principe d'explicabilité de l'IA et son application dans les organisations. Réseaux : communication, technologie, société **N° 240**(4), 179–210 (Sep 2023). https://doi.org/10.3917/res.240.0179, https://hal.science/hal-04225393
50. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law and Technology **31**(2), 841–887 (2018)
51. Wang, X., Yin, M.: Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In: Proceedings of the 26th International Conference on Intelligent User Interfaces. p. 318–328. IUI '21, Association for Computing Machinery (2021)
52. Wiggerthale, J., Reich, C.: Explainable machine learning in critical decision systems: ensuring safe application and correctness. AI **5**(4), 2864–2896 (2024)
53. Williams, O.: Towards human-centred explainable ai: A systematic literature review. Master's Thesis (2021)