MAGIX: a Unified Framework for the Use of XAI in Enterprises

Francesco Sovrano[1,2*] and Tim Miller[3]

[1*]Collegium Helveticum, ETH Zurich, Zurich, 8006, Switzerland.
[2*]Department of Informatics, University of Zurich, Zurich, 8092, Switzerland.
[3]School of Electrical Engineering and Computer Science, University of Queensland, Brisbane, 4072, Queensland, Australia.

*Corresponding author(s). E-mail(s): fsovrano@ethz.ch;
Contributing authors: timothy.miller@uq.edu.au;

Abstract

This chapter proposes a framework for explainable artificial intelligence (XAI) in enterprises. XAI has long been touted as a way to mitigate several issues in enterprise AI systems: (dis)trust, over- and under-reliance, opaqueness, lack of understanding, lack of compliance, bias and discrimination, and poor human decision-making. We argue that, despite having made progress in some of these areas, such as improving understanding and compliance, progress in others has been stunted due to incorrect assumptions about how AI tools are used and understood. We map out a framework for explainable AI that considers some recent positions on the use of explainable AI, with a focus on decision making in enterprises.

Keywords: Explainable AI, Enterprise AI, Reliance, Machine-in-the-loop

This is an Accepted Manuscript version of the following work: Shazia Sadiq (ed.), *Enterprise AI*, 2025, Springer. This version of the manuscript has been accepted for publication, after final editorial and peer review (where applicable) is complete, but is not the Version of Record and does not reflect post-acceptance improvements (such as copyediting or typesetting), or any corrections. The final authenticated version is available online at: `http://dx.doi.org/10.1007/978-3-032-01940-0_6`. Use of this Accepted Manuscript version is subject to the publisher's Accepted Manuscript terms of use: https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms

1 Introduction

The 2023 artificial intelligence (AI) Index Report by Stanford University[1] reveals that, in 2022, global private investment in AI totaled $91.9 billion. The top 11 sectors that received the highest AI private investments were: medical and healthcare ($6.1 billion); data management, processing, and cloud ($5.9 billion); fintech ($5.5 billion); cybersecurity, data protection ($5.4 billion); retail ($4.2 billion); industrial automation, network ($3.9 billion); sales enablement ($3.2 billion); marketing, digital ads ($3.1 billion); AR/VR ($2.4 billion); drones ($1.9 billion); insurtech ($1.7 billion).

This resulted in the integration of AI into several applications across all industry sectors. These AI systems either operate in complete autonomy or by interfacing with human operation through advanced interfaces (Xu et al. 2021).

A significant challenge with advanced AI systems is their "black-box" nature, which makes it hard to understand how they make decisions (Bernal and Mazo 2022). This challenge is critical in almost all fields where AI systems must be accurate, transparent, and accountable to enhance trust in their users and enforce compliance with the law (Sovrano et al. 2021). This is where explainable artificial intelligence (XAI) plays a crucial role, offering tools to make the inner workings of these complex systems more understandable and trustworthy to the diverse stakeholders involved in their operation (Sovrano and Vitali 2024).

The field of XAI provides various methods to interpret machine learning models, differing in strategies, formats, and applicability (Islam et al. 2022). Methods are generally divided into *ante-hoc* models, which are 'inherently interpretable' like decision trees, and *post-hoc* methods that explain black-box models after training. Post-hoc techniques include feature attribution methods assigning importance scores to input features (e.g., LIME (Ribeiro et al. 2016a), SHAP (Lundberg and Lee 2017)); rule-based approaches generating human-understandable rules (e.g., Anchors (Ribeiro et al. 2018)); counterfactual explanations identifying minimal input changes that alter predictions (Guidotti 2024); similarity-based methods comparing inputs to similar instances (Poché et al. 2023); surrogate models approximating complex models with interpretable ones (Bastani et al. 2017); and concept-based methods interpreting the model's internal state using human-friendly concepts (Kim et al. 2018). Explanations also vary in *scope*—global (entire model) vs. local (individual predictions)—and *format*, which can be numerical, textual, or visual.

The main goals of XAI are to show (Fok and Weld 2024):

- How an AI system works from its *inputs*, to avoid *misuse*.
- Whether to trust the AI *outputs* (i.e., whether they are good), to avoid *inappropriate reliance*.

However, evidence indicates that XAI overall does not always mitigate inappropriate reliance (distrust in accurate AI outputs, or excessive trust in AI) or misuse; users often make poorer decisions when presented with incorrect or suboptimal AI suggestions, regardless of XAI (Buçinca et al. 2021; Miller 2023; Vaccaro et al. 2024; Sivaraman et al. 2023). Effectively, the issue with using AI recommendations for decision support is that, when a human decision maker disagrees with the AI recommendation, they struggle to answer: "should I trust the AI

[1] https://aiindex.stanford.edu/report/

or myself?". Empirical evidence suggests that explainability and interpretability, depending on how they are used, may offer limited information to answer this question (Miller 2023; Vaccaro et al. 2024; Siu et al. 2024; Poursabzi-Sangdeh et al. 2021).

Given the variety of XAI methods and the problems stemming from their misuse, which XAI technologies should be used in an enterprise, and when? This chapter addresses this question by providing an in-depth analysis of the potential and actual use of XAI in enterprises.

Taxonomies aid in selecting appropriate methods for different use cases; for instance, Liao et al. (2020) categorize them based on addressed questions, while Vilone and Longo (2021) classify them by stage, scope, and format.

However, current theories do not clearly explain what XAI algorithm to use and when. An early effort to address this gap is presented by Miller (2023), who introduced the *evaluative AI* conceptual framework for explainable decision support. This framework posits that decision aids (i.e., XAI outputs) can reduce inappropriate reliance on AI systems only if they assist users in *verifying and evaluating their own hypotheses about the AI*, rather than simply explaining how the system arrived at its decisions.

Another theoretical perspective is offered by Chen et al. (2023), who argue that machine-generated explanations are only effective when they align with human intuitions. Thus, to determine which XAI algorithm to use in a given context, it is necessary first to identify the wide range of potential human intuitions and hypotheses that might influence decision-making.

To address this challenge, Sovrano and Vilone (2024) proposed framing explanations as answers to questions, where the questions themselves are tied to specific explanatory goals. In their case, these goals were defined by legal requirements, drawing on several European regulations to narrow the scope of possible human intuitions and hypotheses. This approach provided clearer guidance on selecting the appropriate XAI algorithm for specific use cases. While this methodology may seem restrictive, legal frameworks are designed to foster trust and encourage appropriate use of AI by tailoring explainability to user needs. Therefore, aligning XAI goals with legal requirements can, in many cases, coincide with the needs of AI system users. Nonetheless, the methodology proposed by Sovrano and Vilone (2024) is not sufficiently generic to address all possible correct applications of XAI. This limitation underscores the ongoing need for more comprehensive frameworks to guide the design and application of explainable AI systems.

Instead, in this chapter, we present and discuss a novel theoretical framework, *Markovian Action-Observation Generalized Introspection for eXplainability (MAGIX)*, which aims to guide the selection and fitness evaluation of XAI technologies in real-world applications. MAGIX unifies the frameworks proposed by Miller (2023) and Chen et al. (2023) by integrating the XAI-driven hypothesis verification process as a core component of a *machine-in-the-loop Markov Decision Process (MDP)*, a model for sequential decision-making under uncertainty.

Within this *machine-in-the-loop MDP*, MAGIX further organizes human intuitions into three key types of knowledge. This categorization facilitates a mapping function that connects human objectives and hypotheses to relevant human intuitions. Throughout this chapter, we will show how this mapping can help choose the most suitable XAI algorithm for a wide range of realistic enterprise settings (spanning from fintech to healthcare), while also analyzing empirical results from existing literature.

2 Challenges of Current XAI Approaches: The Inappropriate Reliance Issue

Many XAI systems assume users will analytically evaluate explanations to judge AI suggestions. However, this assumption is flawed, as such evaluations require effort, which people often avoid (David et al. 2024). For instance, consider a medical diagnosis system where an AI suggests a diagnosis along with a visual explanation (e.g., heatmaps on medical images). Rather than scrutinizing whether the highlighted regions truly justify the diagnosis, a physician might assume that the explanation indicates the system's overall competence. This "trust by default" leads to inappropriate reliance, especially in high-stakes scenarios.

To address this, Buçinca et al. (2021) developed *cognitive forcing interventions* that require users to engage more deeply with AI explanations. For example, in a financial advisory tool, users might be prompted to explain in their own words why a particular investment recommendation was made. This intervention forces a moment of reflection and helps prevent blind trust. However, such measures have been found to benefit individuals with a high need for cognition (NFC) more than those with low NFC (Cacioppo and Petty 1982). In practice, a risk analyst with high NFC might relish the opportunity to break down the rationale behind an AI's market prediction, whereas an analyst with low NFC may simply bypass or provide minimal engagement with the prompt.

While cognitive forcing interventions aim to reduce inappropriate reliance on AI systems, studies by Vasconcelos et al. (2023) and Fok and Weld (2024) demonstrate that AI explanations can also mitigate this issue during decision-making. Research by Vasconcelos et al. (2023) provides further insight: individuals tend to strategically choose when to engage with AI explanations based on a cost-benefit analysis. For example, in an autonomous driving support system, if the cost (in terms of time and cognitive effort) of verifying why the AI recommended a certain maneuver is low compared to the risk of an accident, a driver may carefully review the explanation. Conversely, if verifying the explanation requires too much effort during a high-pressure situation, the driver may rely on the AI's recommendation without critical examination. This dynamic shows that the design of explainability systems must account for the varying costs users are willing to incur to verify an explanation.

In contrast, Fok and Weld (2024) demonstrate that AI explanations can also assist in decision-making by enabling users to *verify* AI recommendations. Consider a legal document analysis tool where an AI flags potential contractual risks and provides supporting evidence. In this scenario, the explanation serves as a means for legal professionals to efficiently verify the AI's suggestion rather than accepting it at face value. This verification process is analogous to checking a solution in NP-complete problems: while finding a solution can be computationally intensive, verifying a given solution is relatively efficient. Thus, when AI systems present information that facilitates quick verification, they help users avoid the pitfall of inappropriate reliance, ultimately improving decision-making outcomes (Bansal et al. 2021).

This issue of inappropriate reliance can also be understood through dual-process theory (Evans and Stanovich 2013). Humans primarily rely on System 1 thinking, using heuristics and shortcuts for quick decisions. For instance, in a consumer product recommendation system, users might quickly accept AI suggestions based on the aesthetic appeal of the explanation (e.g., attractive charts or simple language) without delving into the underlying data. Although

System 1 is efficient and often effective, its heuristic shortcuts can lead to cognitive biases—resulting in suboptimal outcomes in complex scenarios. Conversely, System 2 thinking is more deliberate and analytical, as might be required in high-stakes contexts like financial risk assessment or critical medical decisions. Yet, due to the significant cognitive effort required, System 2 is engaged less frequently. Even experts, who generally rely on System 1 heuristics built from experience, can be misled when the AI explanation is superficially compelling but substantively shallow. This underscores the challenge: designing interfaces that can encourage a balanced engagement of both systems, especially under varying levels of cognitive motivation.

3 The MAGIX Theoretical Framework

Why are XAI algorithms that answer "why is this decision (in)correct?" (thus supporting decision verification) sometimes more effective at reducing inappropriate reliance than those that explain "how does the input impact the output?"—e.g., SHAP (Lundberg and Lee 2017)—and vice versa? This remains an open question that goes beyond the scope of both evaluative AI (Miller 2023) and the theory of human intuitions proposed by Chen et al. (2023).

The lack of clarity in this area highlights a significant gap in our understanding of how to select XAI algorithms that ensure explanations are sufficiently effective. Addressing this issue is particularly challenging given the combinatorial explosion of human intuitions and hypotheses that XAI systems must meet. Indeed, the quality of an XAI output depends on the user's intended purpose and understanding of the probed AI system.

To address this problem, we introduce a generic theoretical framework inspired by Miller (2023) and Chen et al. (2023). By framing a machine-in-the-loop decision-making process as a special kind of MDP that embeds AI outputs and explanations, we simplify the challenge of defining system and user goals. This model abstracts the possible kinds of intuitions and hypotheses in such a way that significantly reduces the number of their combinations in terms of explanatory needs.

The MAGIX theoretical framework is based on two main steps:

1. The modeling of the use of XAI as part of a machine-in-the-loop MDP.
2. The definition of an action-observation matrix and the corresponding space for informing XAI selection.

3.1 Markov Decision Processes

First, we will provide the essential background information needed to understand what *Markov Decision Processs* are, as these form the basis of the MAGIX framework.

A MDP is a mathematical framework designed to model sequential decision-making scenarios where a human or an agent must choose actions based on the current situation and its expected outcomes. Further, the outcomes of actions can be uncertain, modeled probabilistically. An agent does not know which of the possible outcomes will occur until after the action is executed. Some actions give *rewards*, which can be positive or negative. MDPs can model a standard machine-in-the-loop decision-making scenario.

A MDP captures the interaction between the decision-maker and the environment through a sequence of states, actions, rewards, and transitions, allowing for the modeling of complex

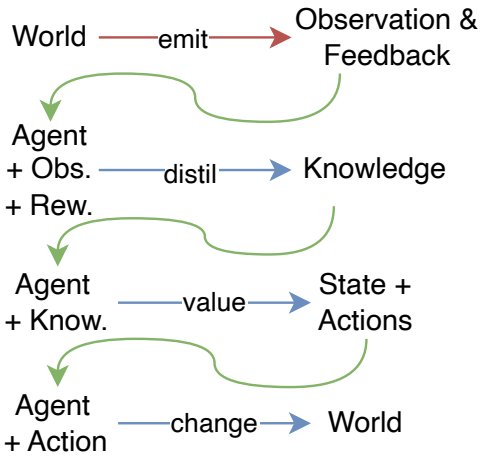


Fig. 1 An overview of a Markov Decision Process. The world emits observations and feedback to an agent, who distills this into knowledge. Using this knowledge, the agent estimates the values of states and actions, and from this, selects an action to execute, which in turn, changes the world.

systems. In our model, the *agent* is the human decision-maker. As shown in Figure 1, the human decision-maker observes the current state, selects an action, and transitions to a new state based on probabilistic dynamics of the system. With every transition, they receive feedback in the form of a reward, which informs their decision-making strategy. The MDP assumes the Markov property, which posits that the future is conditionally independent of the past, given the current state and action. This simplifies the analysis and implementation, making the model a robust choice for real-world problems where decisions unfold over time.

In the context of an MDP, policy functions and value functions are central to decision-making and evaluation.

A *policy function* defines the decision-making strategy, mapping states to actions. It serves as the "plan" the decision-maker (or agent) follows to achieve desired outcomes. In the current state, the policy says which action to take. Once we reach the next state, we use the policy again. A policy can be deterministic, selecting a specific action for each state, or stochastic, assigning probabilities to actions based on the current state.

A *value function*, on the other hand, evaluates the long-term utility of following a policy. It quantifies the expected cumulative reward and comes in two forms: state value functions and action value functions. The *state value function* assesses the expected cumulative reward when starting in a given state and following a specific policy thereafter. It reflects how "desirable" it is to be in a particular state based on future rewards. The *action value function* (often referred to as the Q-function) extends this concept by evaluating the expected cumulative reward of taking a specific action in a given state and then following the policy. This distinction between state and action value functions provides a granular framework for understanding the implications of both being in certain states and making particular choices.

Together, policy and value functions connect directly to the core of an MDP. The policy leverages the insights from value functions to guide decision-making, while the value functions

evaluate the performance of a given policy. This interplay enables the optimization of decision-making strategies, ensuring that the agent makes choices that maximize long-term rewards, even in uncertain environments.

3.2 A Machine-in-the-Loop MDP for XAI

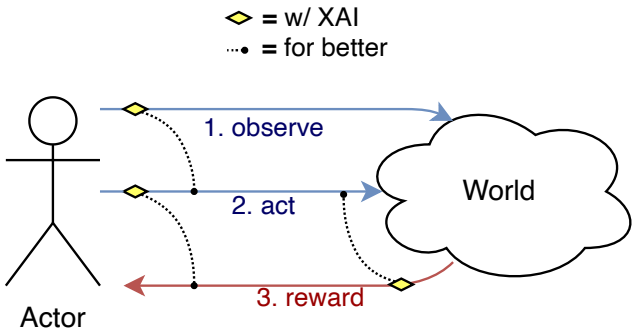


Fig. 2 A MDP enhanced with AI, which assists the actor by improving observation capabilities to achieve better actions, and by recommending "optimal" actions to achieve better rewards.

One compelling example of XAI aiding human decision-making is IBM Research's AIX360 demonstration of an AI-based credit approval system (Research 2020). This system illustrates how different XAI techniques can address the needs of various stakeholders (data scientists, loan officers, and bank consumers) each playing a distinct role in the credit approval process.

In the system, a bank customer submits a loan application, and the bank's AI predicts the risk level associated with the loan, using available customer data. A bank officer reviews this AI-generated risk assessment and makes the final decision on the application. Before deployment, data scientists must validate the AI model's functionality. Meanwhile, customers, as mandated by EU or US laws, are entitled to an explanation of the decision made on their loan application. Each stakeholder benefits from tailored XAI techniques, ensuring transparency, accountability, and actionable insights.

For data scientists, the Generalized Linear Rule Model (GLRM) algorithm (Wei et al. 2019) would be a suitable model. This rule extraction technique helps answer critical questions such as: What is the overall decision-making logic of the model? Is this logic reasonable enough to justify deploying the model? By providing a comprehensive understanding of the AI's internal logic, GLRM fosters confidence in its deployment.

For loan officers, the ProtoDash method (Gurumoorthy et al. 2017) provides example-based explanations that support the interpretation of AI-driven recommendations. It helps answer key questions such as: "Why is this applicant's credit being approved or denied?" and "How can similar past cases inform my decision to grant or reject credit?" By comparing individual predictions with similar historical cases, ProtoDash provides clear, example-based insights that enable loan officers to make more informed and confident decisions.

For bank customers, the Contrastive Explanation Method (CEM) (Dhurandhar et al. 2018) could be used. CEM explains decisions by addressing questions such as: "What changes can

I make to increase the likelihood of approval in future?" This method identifies actionable features a customer can improve, such as reducing the time since the last credit inquiry or increasing the average age of accounts. Furthermore, it highlights potential improvements and their potential impact on the approval outcome.

Together, these XAI techniques create a comprehensive, stakeholder-centric framework. They ensure that data scientists can refine and validate AI systems, loan officers can interpret recommendations effectively, and customers can receive transparent and actionable feedback on their applications. Each of these XAI solutions takes under consideration their user's intuitions and hypotheses, fitting well both the frameworks of Miller (2023) and Chen et al. (2023).

What unifies these diverse uses of XAI? Why are different techniques used, and what criteria guide their selection?

In all cases, stakeholders engage with the "world" to make sense of it. For data scientists, the "world" is the dataset used to train the AI model, and the AI's code-base and logic. For loan officers, it includes customer characteristics and AI-generated assessments. For customers, it encompasses the AI's decision and its underlying rationale. Stakeholders observe the "world", interact with it to achieve specific goals, and receive feedback (implicitly or explicitly) about the outcomes of their actions. This cyclical process of observation, action, and feedback aligns with the framework of a MDP, where the actor aims to maximize cumulative rewards associated with achieving a specific goal.

In this MDP, the actor is augmented by XAI, which aids in properly observing the "world" and helps to explore the actions that can help the actor to achieve their goals. The role of the AI system within the MDP depends on the actor. For some, the AI system may act as a guide, providing personalized recommendations. For others, it may act as part of the "world", influencing the distribution of rewards and shaping how goals are attained. This interplay is encapsulated in the concept of a machine-in-the-loop MDP, as illustrated in Figure 2.

In this model, AI benefits the actor only if it contributes to one or more of the following objectives: (i) improving the observation of the world—step 1 of the MDP, see Figure 2; (ii) enhancing decision-making prior to action execution—step 2; (iii) optimizing outcomes post-action execution—step 3. Each of these objectives corresponds to one of three high-level hypotheses that the actor may aim to verify:

- **H1:** The AI output enhances world observation.
- **H2:** The AI output aids in better decision-making, before acting.
- **H3:** The AI system can be used to achieve the desired objective.

Consequently, the machine-in-the-loop MDP framework organizes the potential uses of XAI into three cases, one per hypothesis, to provide insights about: (i) the actionability of specific knowledge about the world; (ii) the validity of specific actions before they are taken; (iii) the use of the AI system to achieve a good outcome with the action.

Therefore, as summarized in Table 1, in **H1** and **H2** the AI system functions as a recommender system, providing suggestions accompanied by explanations that validate the recommendations. In these scenarios, XAI plays a key role in guiding the actor on how to observe and act within the "world". These cases focus on fostering appropriate reliance on the AI system.

Table 1 Summary of AI Use, Hypotheses, and Goals of XAI. The table outlines how AI can be used to meet key objectives, the hypotheses to be tested, and the corresponding goals of XAI in each scenario.

Use of AI	User Hypotheses	Goals of XAI
AI functions as a recommender system to enhance world observation (Step 1 of the MDP).	**H1:** The AI output enhances world observation.	• Provide insights about the actionability of specific knowledge about the world. • Foster appropriate reliance on the AI system.
AI functions as a recommender system to aid in better decision-making prior to action execution (Step 2 of the MDP).	**H2:** The AI output aids in better decision-making before acting.	• Provide insights about the validity of specific actions before they are taken. • Foster appropriate reliance on the AI system.
AI operates as an autonomous system, making decisions that directly impact the user (Step 3 of the MDP).	**H3:** The AI system can be used to achieve the desired objective.	• Provide insights about the use of the AI system for a good outcome. • Ensure appropriate use of the system.

In contrast, in **H3**, the AI operates as an autonomous system, making decisions that directly impact the user. The explanations in this context address the rationale behind the AI's decisions, as the system executes actions without the actor's ability to directly intervene. This scenario emphasizes the importance of XAI in ensuring the appropriate use of the system.

To illustrate these applications, consider the example of the credit approval system above. Each stakeholder uses XAI at different stages of their respective MDP. Data scientists use XAI to observe the "world" and gain actionable insights for debugging and evaluation (case 1; **H1**). Loan officers rely on XAI to determine the best course of action for handling applications (case 2; **H2**). Customers leverage XAI to understand the system's decisions (i.e., the reward) and adapt their actions to improve outcomes (case 3; **H3**). Through this lens, XAI serves as a critical tool for aligning stakeholders' goals with the broader system's functionality, ensuring more profitable interactions.

Now, we can see an initial answer to the question posed at the beginning of this section: Why does verification support better decision-making than understanding the AI model? In cases 1 and 2, verification helps the decision maker understand their world. In the latter case, instead, standard explainability techniques help the decision maker to understand the AI model itself, which is though only a loose proxy of the world.

Here, we ask the reader to imagine a situation where they need to make an important decision at work, and they received a recommendation from an AI model that they disagree with. What is more useful: *(a)* a reason why the model arrived at that decision (here the 'world' that the XAI deals with is the model); *(b)* a reason why the model recommendation is correct (which may be different to why it made the decision; here the 'world' is the underlying problem space); or *(c)* a reason why the decision-maker's decision is incorrect (here the 'world' is the underlying problem space as well)? We consider option *(a)* to be 'standard' explainable AI, option *(b)* to be the idea of verification such as that defined by Fok and Weld (2024) and Vasconcelos et al. (2023), and option *(c)* to be the idea of evaluative AI (Miller 2023).

3.3 The Action-Observation Matrix

The machine-in-the-loop MDP model aids in narrowing down the possible hypotheses a user might consider. XAI tools are specifically designed to assist users in verifying these hypotheses within the context of a standard sequential decision-making framework. However, what ensures the verification of the goodness of both knowledge and actions? The answer to this question hinges on the user's intuitions regarding the MDP (Chen et al. 2023).

To address this, it is essential to first determine the kind of MDPs in which employing XAI for explaining AI is meaningful. Such scenarios arise when the user's missing knowledge is not explicitly available beforehand (i.e., in the source code of the AI) and must be extracted from a black-box AI model. Specifically, these scenarios arise when there is no guarantee that the AI is error-free due to its reliance on heuristics. This challenge is particularly prominent in:

- **NP-hard problems:** Problems for which no efficient solution exists, making it computationally infeasible to verify the correctness of outcomes in all instances.
- **Ill-defined problems:** Problems where the objectives, constraints, or solution criteria are ambiguous or poorly specified, often necessitating the use of machine learning techniques.

In scenarios outside of these contexts, it is often feasible to build recommender systems that are correct all the time (at least, within their model of the world), rendering the use of XAI for explaining correctness unnecessary, e.g., a symbolic tool that uses probabilities and statistics to calculate risks. Despite this, people may still use less specific, not-always-correct AI solutions to build their applications due to ease, hype, or unawareness of better techniques.

What XAI can effectively explain, instead, are the heuristics underpinning the evaluation of actions and states, as well as the strategies used to extract actionable knowledge from state observations. By explaining these aspects, XAI bridges gaps in users' intuitions, enabling a deeper understanding of the decision-making process finalized to maximizing cumulative rewards.

As discussed in Section 3.2, XAI can be employed to verify three distinct types of hypotheses. Each hypothesis may rely on intuitions regarding the following aspects:

- **I1:** *Knowledge Actionability*—How is actionable knowledge derived? This involves understanding the heuristics that underlie the policy function (see Section 3.1).
- **I2:** *Action Value*—How is the quality of an action evaluated? This relates to the heuristics of the action value function (see Section 3.1).
- **I3:** *State Value*—How is the quality of a state determined? This pertains to the heuristics of the state value function (see Section 3.1).

To verify hypotheses **H1** and **H3**, users may rely on (or lack) intuitions about **I1** and **I3**, as these relate to understanding knowledge actionability and evaluating the value of a state. Conversely, verifying hypothesis **H2** may depend on the user's intuitions (or lack thereof) regarding **I1** and **I2**, as these pertain to evaluating a specific action.

Among these, **I1** is central to all hypotheses because every AI system, by definition, operates on the basis of inputs and outputs, and XAI aims to clarify the relationships between

them. From an XAI standpoint, the heuristics underlying **I1** can either be known or unknown, and the implications vary accordingly:

- If the heuristics of **I1** are known: The XAI tool needs to explain only the execution trace of the heuristics, i.e., what triggered them in the input. This trigger is always initially unknown. It will be enough for the user to verify the correctness of the recommendation/decision.
- If the heuristics of **I1** are unknown: The XAI tool must also explain the heuristics themselves. This involves determining their relevance to the reward function. For example, to verify usability of an AI system (**H3**) the user must know why the state being evaluated is (not) good, which could be explained by showing how to get the desired output from the input.

How the heuristics and their traces are explained depends on the knowledge of the heuristics underlying **I2** and **I3**.

When the rules governing state/action evaluation (heuristics of **I2** and **I3**) are unknown to the user, explaining the heuristics and their trace as explicit rules would be useful. Conversely, if these rules are known, the explanations can just rely on similarity-based reasoning, such as comparing with successful solutions from training data, employing case-based explanations. In these case the user will know how to evaluate the state/action.

This framework reduces the broad range of possible intuitions into two well-defined dimensions per hypothesis, forming the axes of our *action-observation matrix*, as illustrated in Figure 3. Each quadrant in that matrix corresponds to a specific *knowledge acquisition task*. When only knowledge actionability is known, the task is one of *identification or lookup* (Quadrant I). When both knowledge actionability and action/state evaluation are known, the task becomes one of *comparison or browsing* (Quadrant II). Conversely, when neither is known, the task shifts to *clarification or location* (Quadrant III). Finally, in cases where only state/action evaluation is known, the task involves *contextualization or exploration* (Quadrant IV).

For the knowledge acquisition tasks of "Identify" and "Compare", the minimal XAI setting should provide explanations about how the AI's output was generated. Such explanations can be obtained using feature-importance-based XAI algorithms like SHAP (Lundberg and Lee 2017) or LIME (Ribeiro et al. 2016a). For instance, consider an AI system designed to detect cars in an image. Here, SHAP or LIME can explain which features (e.g., shapes, edges, or colors) influenced the detection decision. There is no need to explain what makes a patch of image a car (i.e., the why), the user knows this.

Similarly, when a data scientist needs to debug a credit approval system, they may already understand how to evaluate the rules but require clarity on how input features are connected to the output. Another example involves answering a simple question, such as "When was Einstein born?" For a retrieval-augmented generative AI, an XAI tool like SHAP might highlight the relevance of the terms "Einstein" and "when" in the input, which appear also in the retrieved documents. Without retrieval-augmented generation, however, other generative AI models may lack explainability in such cases.

For tasks involving "Clarify" and "Put in Context", the XAI must provide not only feature attributions but also the reasoning behind the AI's output. This aligns with evaluative AI

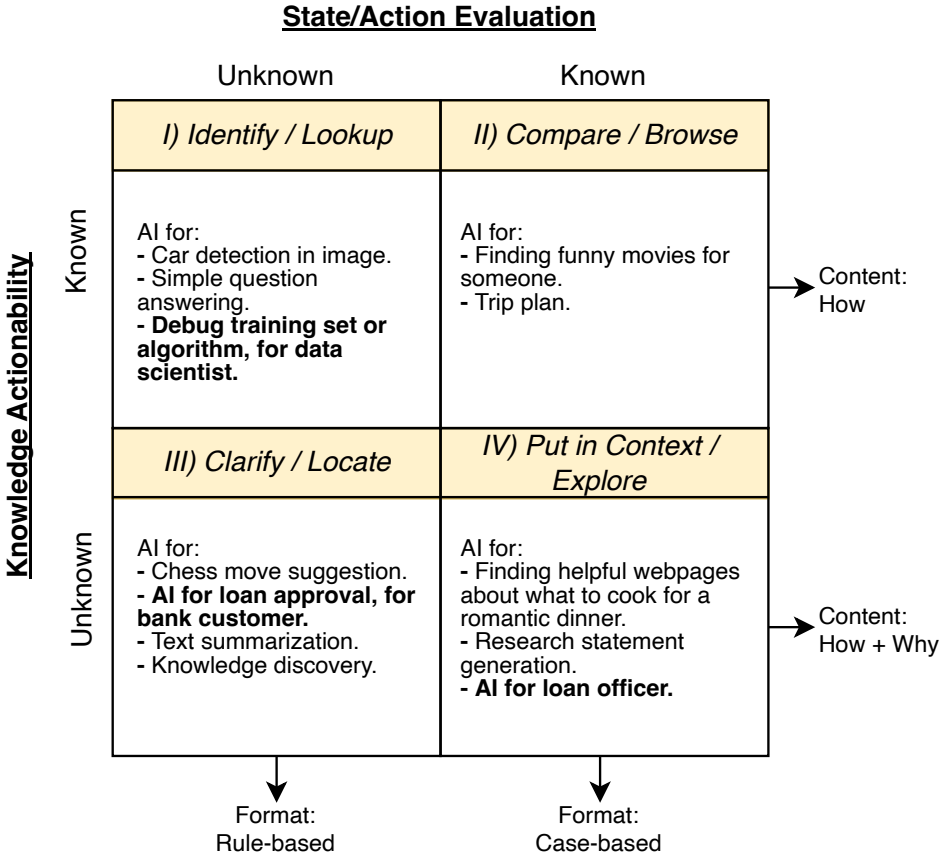


Fig. 3 The action-observation matrix with examples of AI scenarios for each knowledge acquisition task. In bold are those AI scenarios related to the credit approval system used as driving example for this chapter.

(Miller 2023) and findings from Vasconcelos et al. (2023) and Fok and Weld (2024), who emphasize that XAI is most effective when it assists users in verifying the correctness of decisions.

For instance, in the case of a chess move suggestion, the XAI should justify why the suggested move is strategically sound within the broader game context, such as avoiding a checkmate or positioning for a favorable endgame. Similarly, for a cancer detection AI identifying a carcinoma in an image, the XAI should explain why a specific image patch was flagged. This might involve linking the highlighted region to known pathological markers or patterns consistent with carcinoma. In the case of a loan application rejection by a bank's AI, counterfactual explanations are considered as a suitable way to explain model decisions (Wachter et al. 2017). The XAI should indicate how the customer's financial profile could change to meet approval criteria (such as needing a higher income or a lower debt-to-income ratio), providing actionable insights for the applicant.

The distinction between rule-based and case-based formats underlines the flexibility required for different scenarios. For "Identify" and "Clarify" tasks, when the rules for evaluating actions/states are unknown, the explanations are rule-based. For example, in debugging a credit approval system or explaining a straightforward AI decision like recognizing a car in an image, the XAI relies on established algorithms or decision boundaries to communicate how and why the decision was made. In contrast, when the rules are known, as in "Put in Context" or "Explore" tasks, the most minimalist format shifts to case-based explanations. Here, the XAI derives insights from similar cases or analogous scenarios.

However, it is important to also highlight that the concepts of "knowledge actionability" and "action evaluation" are not binary in nature. Rather than being simple states of "known" or "unknown", they exist on a spectrum, with varying degrees of understanding and application across different scales (e.g., global or local). This perspective introduces significant nuances to the concept of "unknown", which may not be fully known in advance.

4 XAI Tools and Technologies

XAI can be viewed both as *approaches* aimed at helping people to understand and appropriately use AI systems, and as *tools* implementing such methods to meet different explanatory needs. In the literature, there is a vast array of domain-dependent and context-specific XAI methods that differ in explanation generation strategies, formats, and applicability to disparate data and learning algorithms (Islam et al. 2022). Several taxonomies guide practitioners in choosing suitable XAI methods for specific tasks. In particular, Adadi and Berrada (2018) offer a comprehensive survey of existing XAI approaches, identifying trends and challenges; van der Velden et al. (2022) focus on XAI for deep learning-based medical imaging, proposing a classification framework; and Nauta et al. (2023) detail 12 conceptual properties for evaluating explanations in machine learning.

One common way to categorize XAI methods is by *stage of application*: *ante-hoc* vs. *post-hoc*. Ante-hoc techniques build inherently interpretable models, whereas post-hoc techniques explain a trained model using an external explainer. Another key dimension is *scope*: methods can provide global (the model's overall logic) or local (explanations of individual predictions) insights (Vilone and Longo 2021). Both types can support local or global understanding; that is, global XAI methods can give insight into local decisions, and a series of local explanations can give a global understanding of an AI model. XAI techniques also vary in *format*: they can be saliency maps or feature-attribution vectors; textual or visual narratives; examples-based statements, rule-based statements; or a combination of these (Liao et al. 2020).

Both ante-hoc and post-hoc XAI methods can serve *ex-ante* explanations (before a decision is taken) or *ex-post* explanations (after a decision). That said, *ex-ante* explanations typically convey global information about the model, while *ex-post* methods often shed light on local, instance-specific behavior. Our classification of XAI methods also considers *model specificity* (model-agnostic vs. model-specific).

Tables 2–3 summarize popular XAI methods organized by the main "explanatory question" they answer (Sovrano and Vilone 2024), their scope, and example references. We highlight how some methods answer "what-if" questions, e.g., counterfactuals (He et al. 2023), while others answer "why" or "how" queries, e.g., saliency maps (Poché et al. 2023), local feature attribution (Ribeiro et al. 2016a), or global feature attribution (Lundberg et al. 2020).

Table 2 Model-Specific XAI Methods with Explanatory Goals.

Question	XAI Algorithms	Applicable Model Types	Explanation
What are the most important features influencing all predictions of the model?	Global feature attribution methods: TreeExplainer (Lundberg et al. 2020), CAVs (Kim et al. 2018)	Tree-based models, Deep Neural Networks (DNNs)	Identify key features for predictions to understand system's conditions of use, risks, and detect anomalies.
How do interactions between features affect all predictions of the model?	Global feature attribution methods (Lundberg et al. 2020)	Tree-based models, Bayesian networks	Identify key feature interactions for system usage, risk, and logic.
What is the model's inner logic?	ExtractRule (Fung et al. 2005)	DNNs, Support Vector Machines (SVMs)	Transparent models mimicking a black-box.
What set of rules does the model usually follow to make all predictions?	Rule-based algorithms (Dash et al. 2018; Wang et al. 2021)	Decision trees, random forests, linear models, DNNs	Clarify the rules for usage, risks, decisions, outputs, or anomalies.
What parts of an input (e.g., an image) influence the model's decision?	Saliency maps (Feldhus et al. 2022), LRP (Bach et al. 2015)	DNN	Identify important input areas to review or contest decisions.
What contributions do individual neurons in a DNN make to the final prediction?	Activation maximization (Li et al. 2022)	DNNs	Analyse neuron contributions for anomaly detection and interpretation.
How do different layers in a DNN contribute to a prediction?	Layer-wise relevance propagation (Li et al. 2022; Kim et al. 2018)	DNNs	Examine layer contributions for anomaly detection, decision logic.
How to interpret a neural net's internal state in terms of human-friendly concepts?	Concept-based methods (Kim et al. 2018)	DNNs	Evaluate key concepts behind decisions, interpret outputs, detect anomalies.

Several XAI algorithms stand out as *particularly influential*:

- *LIME (Local Interpretable Model-Agnostic Explanations)* (Ribeiro et al. 2016b) approximates a black-box model's local behavior around a single instance using a simple, interpretable surrogate (often a sparse linear model). Its main benefit is that it is model-agnostic and easy to implement, offering instance-level feature attributions. However, LIME's reliance on sampling around the input instance assumes local linearity and independence of features, which can be violated in practice; the choice of neighborhood size and sampling distribution can also strongly influence results, risking instability or misleading explanations.

Table 3 Model-Agnostic XAI Methods with Explanatory Goals.

Question	XAI Algorithms	Explanation
What is the inner logic of the model?	Surrogate models (Bastani et al. 2017)	Transparent models mimicking a black-box.
How does a specific feature influence the prediction for an individual instance?	LIME (Ribeiro et al. 2016a), SHAP (Lundberg and Lee 2017)	Assess feature impact on individual outputs to review or interpret decisions.
What are the most important features influencing all predictions of the model?	Global feature attribution methods like SHAP (Lundberg and Lee 2017)	Identify key features for the model's predictions, detect anomalies, assess risk.
What are the most similar instances to a given input with respect to a model's prediction?	Similarity-based methods (Poché et al. 2023)	Compare similar instances to interpret decisions or detect anomalies.
What minimal changes would need to be made to an input to change its prediction?	Counterfactual explanations (Guidotti 2024)	Identify changes to alter outcomes, review or contest decisions, detect anomalies.
Why this output instead of another?	Contrastive explanations (Dhurandhar et al. 2018)	Clarify why one outcome was chosen, interpret or contest decisions, detect anomalies.
What features, when held fixed, are most responsible for a particular prediction or classification?	Anchors (Ribeiro et al. 2018)	Highlights stable feature anchors enabling consistent predictions.
How would changing multiple features simultaneously affect the model's prediction for a specific instance?	Counterfactual and interaction detection methods (Guidotti 2024)	Examine combined feature effects to review or adjust decisions, detect anomalies.
How can we understand the model's decision for a specific instance in the context of its training data?	Contextual analysis methods (Poché et al. 2023)	Provide broader context for local decisions, detect anomalies or biases.

- *SHAP (Shapley Additive Explanations)* (Lundberg and Lee 2017) is grounded in cooperative game theory, assigning each feature an importance score (the Shapley value) indicating its contribution to a prediction. SHAP's appeal lies in its theoretical guarantees of local accuracy, consistency, and additivity, and its extension to various settings (*e.g.*, KernelSHAP for model-agnostic use, TreeSHAP for tree-based models (Lundberg et al. 2020)). Nevertheless, computing Shapley values is expensive for large or complex models; the need to define background (or reference) samples can introduce biases or assumptions about feature independence that do not always hold.
- *Decision tree surrogate models and other rule-based methods* (Von Winterfeldt and Edwards 1993; Dash et al. 2018) inherently yield human-readable rules (via splits or Boolean conditions), making them popular *inherently interpretable* or *surrogate* approaches for larger black-boxes. Their clarity is also their weakness, as complex decision boundaries or feature interactions might be missed or oversimplified, especially when a tree-based surrogate is used for high-dimensional data.
- *Counterfactual explanations* (Mothilal et al. 2020; Guidotti 2024) reveal how to alter an input minimally to obtain a different outcome. They offer actionable insights and local transparency. However, defining actionable or realistic changes demands careful

consideration of domain constraints and feature correlations; if a model's behavior is non-linear, it is difficult to guarantee that minimal changes found in one region are meaningful in another.

- *Anchors* (Ribeiro et al. 2018) highlights a short set of rule-like conditions that anchor (almost guarantee) the same classification outcome. While anchors provide concise, high-precision local explanations, they depend on user-defined perturbation distributions and can become overly specific. Coverage may be low in complex domains, and the logic behind anchor generation can be opaque for non-technical users.
- *Grad-CAM* (Selvaraju et al. 2020) and *Integrated Gradients* (Sundararajan et al. 2017) are representative *gradient-based* explanation methods for deep networks. They create saliency or relevance maps for images or other differentiable inputs, helping to pinpoint which regions influenced the prediction. Though widely used for visual data and medically oriented tasks, these methods can be sensitive to choice of baselines (Integrated Gradients) or susceptible to resolution limits (Grad-CAM), and their reliance on gradients alone can fail when activation saturates or in the presence of adversarial manipulations.
- *Concept-based* explanations, especially for computer vision (Kim et al. 2018), aim to identify human-aligned concepts present in images that contribute to a model's predictions. These methods can be either supervised (Koh et al. 2020), requiring training data annotated with relevant concepts, or unsupervised (Zhang et al. 2021; Ghorbani et al. 2019), which rely only on class labels. Concept-based approaches have been shown to be significantly more helpful to users than saliency-based methods. However, supervised techniques demand extensive and often noisy or costly labeling, while unsupervised methods frequently discover concepts that are difficult for humans to interpret.

In summary, XAI methods encompass a broad spectrum of approaches (model-agnostic vs. model-specific, global vs. local, ante-hoc vs. post-hoc, etc.), each rooted in different assumptions and suited to different "explanatory" questions. For example, TreeSHAP (Lundberg et al. 2020) helps uncover global feature attributions in ensemble tree models; LIME (Ribeiro et al. 2016a) provides local interpretability for any black-box, but can be sensitive to sampling issues; and gradient-based saliency approaches excel at visualizing relevant pixels or features in DNNs, but can struggle with non-differentiable components. None of these methods, however, address downstream consequences of decisions; they focus on clarifying the *process* leading to a prediction or classification. Hence, while XAI tools can greatly aid in explaining or justifying AI outputs, additional manual steps are often required to ensure domain-specific compliance, correctness, and consideration of real-world consequences.

5 Applications of XAI in Enterprises: Case Studies and Analysis

What XAI to use in an enterprise? The answer depends on the action-observation matrix and the types of hypotheses involved. As we illustrated in the previous section, selecting an appropriate XAI algorithm is inherently both intuition-dependent and hypothesis-driven. Together, these two dimensions define the verifiability task.

In this section, we will demonstrate how to apply the MAGIX framework to systematically map AI tasks to XAI applications. Specifically, we will provide one example for each

Table 4 Examples of AI tasks and XAI solutions in the medical and healthcare sector, categorized by quadrant and hypothesis.

Quad.	Hyp.	AI User	AI Task	XAI
I	H1	Radiologist	Identify shards of broken bones	Saliency maps or concept-based explanations to highlight tumor regions in scans
II	H1	Medical Student	Compare AI diagnosis with known cases	Example-based explanations; retrieving similar cases
III	H1	Researcher	Clarify unknown biomarkers in genetic data	Rule-based explanations; explaining why the suggestion is good/valid
IV	H1	Doctor	Explore alternative diagnoses for complex symptoms	Example-based explanations; visual step-by-step procedures
I	H2	Medical Researcher	Identify novel experimental therapies for patient	LIME or SHAP explanations
II	H2	Physician	Compare treatment options suggested by AI	Prototype Selection
III	H2	Doctor	Clarify potential drug interactions before prescribing	Rule-based explanations; displaying interaction rules
IV	H2	Surgeon	Explore new surgical techniques	Example-based explanations; visual step-by-step procedures
I	H3	Patient	Confirm specialized treatment	Feature Attribution (e.g., SHAP)
II	H3	Patient	Impose lifestyle changes	Example-based explanations
III	H3	Patient	Approve/deny treatment coverage	Counterfactual explanations
IV	H3	Patient	Adjust medications	Case-based reasoning

combination of hypothesis type and knowledge acquisition task (i.e., a quadrant in the action-observation; see Figure 3). Our focus will be on the sectors with the highest AI investments, according to Stanford University's 2023 AI Index Report. These sectors include medical and healthcare; data management, processing, and cloud; fintech; cybersecurity and data protection; retail; industrial automation and networks; sales enablement; marketing and digital advertising; augmented and virtual reality (AR/VR); drones; and insurtech.

Since providing 12 examples for each sector would amount to a total of 132 examples, we have chosen to present all the examples for the medical and healthcare sector (see Table 4) and only a few relevant examples for the other sectors (see Table 5).

5.1 Medical and Healthcare Sector

Given an AI system to be explained, the MAGIX method involves the following steps:

1. Identify the Hypothesis (**H1**, **H2**, **H3**): Determine which high-level hypothesis the AI user aims to validate about the system.
2. Determine the Knowledge Acquisition Task: Based on the user's knowledge and actionability, categorize the task into one of the four quadrants of the action-observation matrix.
3. Select Appropriate XAI Methods: choose XAI algorithms that provide the necessary explanations based on the task type.

To construct Table 4, we started by identifying AI applications (and their users) relevant to the medical and healthcare sector, focusing on tasks that align with real-world use cases such as identifying tumors, comparing treatment options, or exploring surgical techniques. Then, we used the MAGIX framework, following the steps above, to understand what XAI algorithms are needed.

For instance, for the AI application of identifying shards of broken bones in CT (Computed Tomography) scans, the users are radiologists who know how to understand whether the AI output is correct (known knowledge actionability) but do not know how to assess the image because it takes too much time and effort on a positive identification or some shards are so small they can be missed (unknown state evaluation). This task fits into Quadrant I, with Hypothesis **H1** since the AI output enhances world observation. An appropriate XAI method is feature attribution (e.g., Grad-CAM or saliency maps), which highlights regions in the CT scans indicative of bone shards.

For the task of comparing AI diagnoses with known cases, the user is a medical student able to understand whether a diagnosis is correct (known knowledge actionability) and to assess the state of a patient (known state evaluation). This task fits into Quadrant II with Hypothesis **H1**, as the AI helps in comparing knowledge. Example-based explanations, such as retrieving similar cases, would be suitable to explain how the AI's diagnosis aligns with existing knowledge, since the medical student does not need to be given the specific rules governing each diagnosis (they are already known).

When clarifying unknown biomarkers in genetic data, the users are researchers who do not have knowledge to evaluate the patterns in genetic sequences (unknown knowledge actionability) and do not have a framework to assess these patterns' significance (unknown state evaluation). If they knew these things, the AI tool would be unnecessary. This task fits into Quadrant III under Hypothesis **H1**. Rule-based explanations are used to translate complex genetic patterns into interpretable if-then statements that make the recommendations accessible and assessable. For example, a rule-based explainer might state: "If a genetic sequence exhibits mutation X and expression level of gene Y is above threshold Z, then it indicates a potential biomarker for disease A". By breaking down intricate genetic data into simple, logical rules, researchers may be able to grasp the significance of specific patterns without needing extensive prior knowledge. This approach not only clarifies how the recommendation was derived but also allows researchers to evaluate its validity and relevance, thereby demonstrating the recommendation's effectiveness.

For the task of exploring alternative diagnoses for complex symptoms, the user is a clinician who is not familiar with possible diagnoses due to the large potential number (unknown knowledge actionability) but knows how to assess them when they see one (known state evaluation). This task belongs to Quadrant IV under Hypothesis **H1**. Example-based explanations, including visual step-by-step procedures, are appropriate to contextualize the AI's suggested diagnoses, showing how they relate to known cases.

Identifying experimental therapies for patients involves a medical researcher who knows how to interpret suggested therapies once they are recommended (known knowledge actionability) but is exploring therapies that are outside their current knowledge base (unknown action evaluation). This task fits Quadrant I under Hypothesis **H2**. Feature attribution methods like LIME or SHAP explain how the AI determined its recommendations.

For comparing treatment options suggested by AI, the user could be a physician who is knowledgeable about treatment possibilities (known knowledge actionability) and knows how to evaluate their outcomes (known action evaluation). This task fits Quadrant II under Hypothesis **H2**. Prototype selection is used to provide representative examples of the treatments, enabling comparisons and ensuring transparency about the AI's suggestions.

Clarifying potential drug interactions before prescribing involves a doctor who is unsure of the rules governing all interactions (unknown knowledge actionability) and does not know how these interactions might affect patient outcomes (unknown action evaluation). This fits into Quadrant III under Hypothesis **H2**. Rule-based explanations are used to display interaction rules, explaining the conditions under which specific drug combinations are unsafe.

For exploring new surgical techniques, the user is a surgeon unsure of how to implement them effectively (unknown knowledge actionability) but aware of their goodness (known action evaluation). This task belongs to Quadrant IV under Hypothesis **H2**. Example-based explanations and visual step-by-step procedures contextualize the AI's suggestions and provide guidance for practical implementation.

For the task of confirming specialized treatment, the user is a patient who trusts the prescribed treatment (known knowledge actionability) but lacks the capacity to evaluate why the AI suggested it (unknown state evaluation). This task fits Quadrant I under Hypothesis **H3**. Feature attribution methods such as SHAP are used to explain how the AI's decision aligns with the patient's medical profile.

In the case of comparing AI instructions for lifestyle changes, the user is a patient familiar with lifestyle adjustment strategies (known knowledge actionability) and capable of evaluating whether they fit their life (known state evaluation). This fits into Quadrant II under Hypothesis **H3**. Example-based explanations illustrate how similar cases benefited from specific lifestyle changes, providing relatable insights.

When clarifying the AI's denial of treatment coverage, the user is a patient who does not understand the system's decision-making rules (unknown knowledge actionability) and is unable to evaluate how to change their profile to meet approval criteria (unknown state evaluation). This task belongs to Quadrant III under Hypothesis **H3**. Counterfactual explanations reveal how changes in the patient's circumstances could lead to approval, providing actionable insights.

Finally, adjusting medications based on AI suggestions involves a patient who does not understand their current medications (unknown knowledge actionability) but due to previous experience is confident about how modifications might affect their health outcomes (known state evaluation). This fits into Quadrant IV under Hypothesis **H3**. Case-based reasoning thus may offer examples of similar cases where adjustments were successful and why, giving the patient context and confidence in their decision.

5.2 Other Sectors

To construct Table 5, we started by identifying important and realistic AI use cases in the specified sectors, ensuring they align with real-world applications such as fraud detection in fintech, intrusion detection in cybersecurity, or personalized product recommendations in retail. For each use case, we determined the AI user and the specific task they aim to accomplish. Then, we used the MAGIX framework to assign the appropriate quadrant based

Table 5 Examples of AI tasks and XAI solutions across various sectors, categorized by sector, quadrant, and hypothesis.

Sector	Quad.	Hyp.	AI User	AI Task	XAI
Data Management	II	H2	Cloud Infrastructure Manager	Compare resource allocation options	Feature attribution methods
Fintech	I	H1	Fraud Analyst	Identify fraudulent transactions	Feature attribution methods
Fintech	II	H2	Loan Officer	Compare loan applications for approval	Prototype selection
Cybersec.	III	H2	Cybersecurity Specialist	Clarify user behavior anomalies	Counterfactual explanations
Retail	II	H3	Customer	Compare personalized product recommendations	Example-based explanations
Industrial Automation	I	H1	Maintenance Engineer	Identify equipment failure indicators	Feature attribution methods
Industrial Automation	II	H2	Process Engineer	Compare robotic workflow optimizations	Prototype selection
Sales Enablement	II	H2	Sales Representative	Compare lead scoring results	Feature attribution methods
Sales Enablement	I	H3	Customer	Understand automated email responses	Feature attribution methods
Marketing	I	H1	Social Media Analyst	Identify brand sentiment indicators	Feature attribution methods
AR/VR	IV	H3	Trainee	Put in context VR training feedback	Example-based explanations
Drones	I	H1	Drone Operator	Identify navigation obstacles	Feature attribution methods
Insurtech	II	H2	Claims Adjuster	Compare insurance claims	Prototype selection

on the user's knowledge and actionability, identified the relevant hypothesis, and selected suitable XAI methods according to the guidelines provided.

This table includes up to two examples per sector, covering 14 AI use cases. Each entry specifies the sector, quadrant, hypothesis, AI user, AI task, and the selected XAI method.

6 Conclusion

In this chapter, we introduced the *MAGIX* theoretical framework to address the critical challenge of selecting appropriate XAI technologies in enterprise settings. By modeling the human-AI interaction as a machine-in-the-loop Markov Decision Process, MAGIX provides a systematic approach to align XAI methods with users' hypotheses and intuitions. This alignment is facilitated through the action-observation matrix, which categorizes knowledge acquisition tasks and guides the selection of suitable XAI algorithms.

Our framework unifies and extends previous theories by integrating the evaluative AI concept and the importance of human intuitions in explanation effectiveness. Through case

studies in the medical and healthcare sector, as well as examples from other industries, we demonstrated how MAGIX can be applied to real-world scenarios, ensuring that XAI tools not only enhance transparency but also effectively support users' decision-making processes.

Future work includes validating the MAGIX framework through empirical studies and expanding it to accommodate emerging AI technologies and use cases. By continuing to refine and test this framework, we aim to contribute to the development of more effective and user-centered XAI systems that meet the diverse needs of stakeholders across various industries.

Acknowledgements

F. Sovrano acknowledges the support of the Swiss National Science Foundation for SNF Project 200021_197227 and of the UZH Global Strategy and Partnerships Funding Scheme.

References

Adadi A, Berrada M (2018) Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052, URL https://doi.org/10.1109/ACCESS.2018.2870052

Bach S, Binder A, Montavon G, et al (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one

Bansal G, Wu T, Zhou J, et al (2021) Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In: Kitamura Y, Quigley A, Isbister K, et al (eds) CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021. ACM, pp 81:1–81:16, https://doi.org/10.1145/3411764.3445717, URL https://doi.org/10.1145/3411764.3445717

Bastani O, Kim C, Bastani H (2017) Interpretability via model extraction. arXiv preprint arXiv:170609773

Bernal J, Mazo C (2022) Transparency of artificial intelligence in healthcare: insights from professionals in computing and healthcare worldwide. Applied Sciences 12(20):10228

Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think: Cognitive forcing functions can reduce overreliance on AI in ai-assisted decision-making. Proc ACM Hum Comput Interact 5(CSCW1):188:1–188:21. https://doi.org/10.1145/3449287, URL https://doi.org/10.1145/3449287

Cacioppo JT, Petty RE (1982) The need for cognition. Journal of personality and social psychology 42(1):116

Chen C, Feng S, Sharma A, et al (2023) Machine explanations and human understanding. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023. ACM, p 1, https://doi.org/10.1145/3593013.3593970, URL https://doi.org/10.1145/3593013.3593970

Dash S, Günlük O, Wei D (2018) Boolean decision rules via column generation. In: Bengio S, Wallach HM, Larochelle H, et al (eds) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pp 4660–4670, URL https://proceedings.neurips.cc/paper/2018/hash/743394beff4b1282ba735e5e3723ed74-Abstract.html

David L, Vassena E, Bijleveld E (2024) The unpleasantness of thinking: A meta-analytic review of the association between mental effort and negative affect. Psychological bulletin https://doi.org/10.1037/bul0000443, URL https://psycnet.apa.org/record/2025-07056-001?doi=1

Dhurandhar A, Chen P, Luss R, et al (2018) Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In: Bengio S, Wallach HM, Larochelle H, et al (eds) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pp 590–601, URL https://proceedings.neurips.cc/paper/2018/hash/c5ff2543b53f4cc0ad3819a36752467b-Abstract.html

Evans JSB, Stanovich KE (2013) Dual-process theories of higher cognition: Advancing the debate. Perspectives on psychological science 8(3):223–241

Feldhus N, Hennig L, Nasert MD, et al (2022) Constructing natural language explanations via saliency map verbalization. arXiv preprint arXiv:221007222

Fok R, Weld DS (2024) In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. AI Mag 45(3):317–332. https://doi.org/10.1002/AAAI.12182, URL https://doi.org/10.1002/aaai.12182

Fung G, Sandilya S, Rao RB (2005) Rule extraction from linear support vector machines. In: Grossman R, Bayardo RJ, Bennett KP (eds) Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005. ACM, pp 32–40, https://doi.org/10.1145/1081870.1081878, URL https://doi.org/10.1145/1081870.1081878

Ghorbani A, Wexler J, Zou JY, et al (2019) Towards automatic concept-based explanations. Advances in neural information processing systems pp 9273–9282. URL https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html

Guidotti R (2024) Counterfactual explanations and how to find them: literature review and benchmarking. Data Min Knowl Discov 38(5):2770–2824. https://doi.org/10.1007/S10618-022-00831-6, URL https://doi.org/10.1007/s10618-022-00831-6

Gurumoorthy KS, Dhurandhar A, Cecchi G (2017) Protodash: Fast interpretable prototype selection. arXiv preprint arXiv:170701212

He M, An B, Wang J, et al (2023) Counterfactual explanations for sequential recommendation with temporal dependencies. In: Zhang F, Wang H, Barhamgi M, et al (eds) Web Information

Systems Engineering - WISE 2023 - 24th International Conference, Melbourne, VIC, Australia, October 25-27, 2023, Proceedings, Lecture Notes in Computer Science, vol 14306. Springer, pp 533–543, https://doi.org/10.1007/978-981-99-7254-8_41, URL https://doi.org/10.1007/978-981-99-7254-8_41

Islam MR, Ahmed MU, Barua S, et al (2022) A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Applied Sciences 12(3). https://doi.org/10.3390/app12031353, URL https://www.mdpi.com/2076-3417/12/3/1353

Kim B, Wattenberg M, Gilmer J, et al (2018) Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Dy JG, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, Proceedings of Machine Learning Research, vol 80. PMLR, pp 2673–2682, URL http://proceedings.mlr.press/v80/kim18d.html

Koh PW, Nguyen T, Tang YS, et al (2020) Concept Bottleneck Models. International Conference on Machine Learning abs/2007.04612:5338–5348. URL https://proceedings.mlr.press/v119/koh20a, 2007.04612

Li J, Zhang C, Zhou JT, et al (2022) Deep-lift: Deep label-specific feature learning for image annotation. IEEE Trans Cybern 52(8):7732–7741. https://doi.org/10.1109/TCYB.2021.3049630, URL https://doi.org/10.1109/TCYB.2021.3049630

Liao QV, Gruen DM, Miller S (2020) Questioning the AI: informing design practices for explainable AI user experiences. In: Bernhaupt R, Mueller FF, Verweij D, et al (eds) CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020. ACM, pp 1–15, https://doi.org/10.1145/3313831.3376590, URL https://doi.org/10.1145/3313831.3376590

Lundberg SM, Lee S (2017) A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, et al (eds) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp 4765–4774, URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Lundberg SM, Erion GG, Chen H, et al (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2(1):56–67. https://doi.org/10.1038/S42256-019-0138-9, URL https://doi.org/10.1038/s42256-019-0138-9

Miller T (2023) Explainable AI is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative AI. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023. ACM, pp 333–342, https://doi.org/10.1145/3593013.3594001, URL https://doi.org/10.1145/3593013.3594001

Mothilal RK, Sharma A, Tan C (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In: Hildebrandt M, Castillo C, Celis LE, et al (eds) FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020. ACM, pp 607–617, https://doi.org/10.1145/3351095.3372850, URL https://doi.org/10.1145/3351095.3372850

Nauta M, Trienes J, Pathak S, et al (2023) From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Comput Surv 55(13s):295:1–295:42. https://doi.org/10.1145/3583558, URL https://doi.org/10.1145/3583558

Poché A, Hervier L, Bakkay MC (2023) Natural example-based explainability: A survey. In: Longo L (ed) Explainable Artificial Intelligence - First World Conference, xAI 2023, Lisbon, Portugal, July 26-28, 2023, Proceedings, Part II, Communications in Computer and Information Science, vol 1902. Springer, pp 24–47, https://doi.org/10.1007/978-3-031-44067-0_2, URL https://doi.org/10.1007/978-3-031-44067-0_2

Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, et al (2021) Manipulating and Measuring Model Interpretability. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, p 1–52, https://doi.org/10.1145/3411764.3445315, URL https://doi.org/10.1145/3411764.3445315

Research I (2020) Ai explainability 360 - demo. https://aix360.res.ibm.com/data, accessed: 2024-11-25

Ribeiro MT, Singh S, Guestrin C (2016a) Model-agnostic interpretability of machine learning. CoRR abs/1606.05386. URL http://arxiv.org/abs/1606.05386, 1606.05386

Ribeiro MT, Singh S, Guestrin C (2016b) "why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram B, Shah M, Smola AJ, et al (eds) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM, pp 1135–1144, https://doi.org/10.1145/2939672.2939778, URL https://doi.org/10.1145/2939672.2939778

Ribeiro MT, Singh S, Guestrin C (2018) Anchors: High-precision model-agnostic explanations. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. AAAI Press, pp 1527–1535, https://doi.org/10.1609/AAAI.V32I1.11491, URL https://doi.org/10.1609/aaai.v32i1.11491

Selvaraju RR, Cogswell M, Das A, et al (2020) Grad-cam: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128(2):336–359. https://doi.org/10.1007/S11263-019-01228-7, URL https://doi.org/10.1007/s11263-019-01228-7

Siu HC, Leahy K, Mann M (2024) How interpretable is "interpretable" machine learning? MIT Case Studies in Social and Ethical Responsibilities of Computing https://doi.org/10.21428/2c646de5.9e0237cb, URL https://mit-serc.pubpub.org/pub/5kduhf9d/download/pdf

Sivaraman V, Bukowski LA, Levin J, et al (2023) Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. In: Schmidt A, Väänänen K, Goyal T, et al (eds) Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023. ACM, pp 754:1–754:18, https://doi.org/10.1145/3544548.3581075, URL https://doi.org/10.1145/3544548.3581075

Sovrano F, Vilone G (2024) Aligning XAI with EU regulations for smart biomedical devices: A methodology for compliance analysis. In: Endriss U, Melo FS, Bach K, et al (eds) ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024), Frontiers in Artificial Intelligence and Applications, vol 392. IOS Press, pp 826–833, https://doi.org/10.3233/FAIA240568, URL https://doi.org/10.3233/FAIA240568

Sovrano F, Vitali F (2024) Explanatory artificial intelligence (YAI): human-centered explanations of explainable AI and complex data. Data Min Knowl Discov 38(5):3141–3168. https://doi.org/10.1007/S10618-022-00872-X, URL https://doi.org/10.1007/s10618-022-00872-x

Sovrano F, Sapienza S, Palmirani M, et al (2021) A survey on methods and metrics for the assessment of explainability under the proposed AI act. In: Schweighofer E (ed) Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021, Frontiers in Artificial Intelligence and Applications, vol 346. IOS Press, pp 235–242, https://doi.org/10.3233/FAIA210342, URL https://doi.org/10.3233/FAIA210342

Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Precup D, Teh YW (eds) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, Proceedings of Machine Learning Research, vol 70. PMLR, pp 3319–3328, URL http://proceedings.mlr.press/v70/sundararajan17a.html

Vaccaro M, Almaatouq A, Malone T (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature human behaviour pp 1–11. https://doi.org/10.1038/s41562-024-02024-1, URL https://www.nature.com/articles/s41562-024-02024-1

Vasconcelos H, Jörke M, Grunde-McLaughlin M, et al (2023) Explanations can reduce overreliance on AI systems during decision-making. Proc ACM Hum Comput Interact 7(CSCW1):1–38. https://doi.org/10.1145/3579605, URL https://doi.org/10.1145/3579605

van der Velden BHM, Kuijf HJ, Gilhuijs KGA, et al (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Anal

79:102470. https://doi.org/10.1016/J.MEDIA.2022.102470, URL https://doi.org/10.1016/j.media.2022.102470

Vilone G, Longo L (2021) Classification of explainable artificial intelligence methods through their output formats. Mach Learn Knowl Extr 3(3):615–661. https://doi.org/10.3390/MAKE3030032, URL https://doi.org/10.3390/make3030032

Von Winterfeldt D, Edwards W (1993) Decision analysis and behavioral research. Cambridge University Press, 1986

Wachter S, Mittelstadt BD, Russell C (2017) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. CoRR abs/1711.00399. URL http://arxiv.org/abs/1711.00399, 1711.00399

Wang J, Wiens J, Lundberg SM (2021) Shapley flow: A graph-based approach to interpreting model predictions. In: Banerjee A, Fukumizu K (eds) The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event, Proceedings of Machine Learning Research, vol 130. PMLR, pp 721–729, URL http://proceedings.mlr.press/v130/wang21b.html

Wei D, Dash S, Gao T, et al (2019) Generalized linear rule models. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97. PMLR, pp 6687–6696, URL http://proceedings.mlr.press/v97/wei19a.html

Xu W, Dainoff MJ, Ge L, et al (2021) From human-computer interaction to human-ai interaction: new challenges and opportunities for enabling human-centered ai. arXiv preprint arXiv:210505424 5

Zhang R, Madumal P, Miller T, et al (2021) Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors. Proceedings of the AAAI Conference on Artificial Intelligence 35(13):11682–11690. https://doi.org/10.1609/aaai.v35i13.17389, URL https://ojs.aaai.org/index.php/AAAI/article/view/17389