ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 9, No. 5, 9-20 (2024)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Advancements in Explainable Artificial Intelligence for Enhanced Transparency and Interpretability across Business Applications

Maikel Leon[*,1], Hanna DeSimone[2]

[1]*Department of Business Technology, Miami Herbert Business School, University of Miami, Miami, Florida, USA*

[2]*Miami Herbert Business School, University of Miami, Miami, Florida, USA*

ARTICLE INFO

Article history:
Received: 23 July, 2024
Revised: 14 September, 2024
Accepted: 15 September, 2024
Online: 20 September, 2024

Keywords:
Explainable AI
Transparency
Interpretability

ABSTRACT

This manuscript offers an in-depth analysis of Explainable Artificial Intelligence (XAI), emphasizing its crucial role in developing transparent and ethically compliant AI systems. It traces AI's evolution from basic algorithms to complex systems capable of autonomous decisions with self-explanation. The paper distinguishes between explainability—making AI decision processes understandable to humans—and interpretability, which provides coherent reasons behind these decisions. We explore advanced explanation methodologies, including feature attribution, example-based methods, and rule extraction technologies, emphasizing their importance in high-stakes domains like healthcare and finance. The study also reviews the current regulatory frameworks governing XAI, assessing their effectiveness in keeping pace with AI innovation and societal expectations. For example, rule extraction from artificial neural networks (ANNs) involves deriving explicit, human-understandable rules from complex models to mimic explainability, thereby making the decision-making process of ANNs transparent and accessible. Concluding, the paper forecasts future directions for XAI research and regulation, advocating for innovative and ethically sound advancements. This work enhances the dialogue on responsible AI and establishes a foundation for future research and policy in XAI.

1. Introduction

Artificial Intelligence (AI) has become an increasingly popular topic in recent years. AI is defined as the capability of a machine to replicate cognitive functions associated with the human mind [1]. As new technologies like ChatGPT emerge, uncertainty about the impact of AI technologies on the business world is steadily growing. The complexity of these systems makes it difficult to understand how AI arrives at its conclusions, resulting in a "black box" scenario where the process used to come to a system output is not fully transparent [2]. The black box syndrome in such systems can create problems in critical fields like finance and medical applications. These fields require more transparency and trust when diagnosing or approving a loan. As the use of AI grows, the demand for explainability within knowledge-based systems increases. In the business community, there is worry about human trust in AI recommendations, leading to a desire for transparency in AI systems [3]. The current lack of transparency in AI systems has led to increased focus on the research of explainable AI. There is a clear need for explainability, trust, and transparency in algorithms across various applications. The concept of Explainable AI generalizes new possibilities for AI programs.

The surge in the adoption of AI systems across various sectors necessitates a parallel increase in explainability to ensure these systems are trustworthy, ethical, and accessible. Here are some concise reasons:

- Regulatory Compliance: Increasing global regulations around data privacy and AI transparency demand mechanisms for explaining and justifying automated decisions, especially in critical sectors like healthcare, finance, and legal.
- Ethical Considerations: As AI systems become more prevalent, the ethical implications of their decisions become more significant. Explainable AI facilitates the understanding of automated decisions, supporting ethical auditing and accountability.
- User Trust: Transparency in AI operations fosters user trust and acceptance, crucial for the widespread deployment of AI technologies in sensitive and impactful areas.

These points underscore the essential role of explainability in the responsible scaling of AI technologies. As we delve deeper into the nuances of AI applications, the complexity of these systems grows [4], highlighting the urgent need for advanced research

[*]Corresponding Author: Maikel Leon (mleon@miami.edu)

in explainable AI. Such research not only aids in aligning AI systems with human values and norms but also opens new avenues for innovation in AI governance and policy-making.

2. AI's history highlights

In recent years, the explainability of systems has emerged as a significant factor in adopting AI. It has become essential for practical, social, and legal reasons that users are provided with an explanation of how a system reaches a particular output [5]. Explanations are necessary to understand a system's functions and give users insight into debugging system issues. However, experts have not defined a reason or the qualities it must possess [5]. In early forms of AI, explainability was not prioritized. The origin of AI can be traced back to the 1940s. These first roots of AI were found in the WWII code-breaking machine developed by English mathematician Alan Turing [6]. The technology's ability to outperform humans in decoding caused Turing to question the system's intelligence. In 1950, Turing released a paper discussing how to produce intelligent systems and test their intelligence. In summary, he proposed a test that considered a machine intelligent if a human cannot distinguish between another human and the machine [6]. Today, the Turing Test is still utilized as a benchmark for recognizing the intelligence of a system.

AI foundation traces back to 1956 at Dartmouth College, which kick-started a new era of machine learning research and development. The first hint of explainability can be found in early knowledge-based expert systems in the 1960s. Rule-based expert systems utilize expert human knowledge to solve problems that usually require human-level intelligence [7]. Using expert or domain knowledge, these software systems assist humans in decision-making. Expert systems use an approach of "if-then" statements and have several essential parts, including a knowledge base (usually formatted as a set of rules), an inference engine, and an interface to convey information to a user [6, 7]. Using a top-down approach, expert systems can quickly formalize human intelligence into logical rules that can be followed step-by-step. 1966, at MIT, Joseph Weizenbaum created ELIZA, a natural language processing tool capable of conversing with a human user [6]. ELIZA was one of the first programs to pass the Turing Test. In the early 1970s, governments began to hesitate and pull back funding for AI research, causing a gap in the development of AI.

In the 1980s, Expert Systems, using AI-derived symbolic reasoning techniques to address complex problems, began demonstrating the technology's ability to achieve a firm's goals [8]. However, critics began to argue that overall, expert systems rarely achieved their set goals and, in many cases, could not achieve expert-level performance [8, 9]. These concerns heavily came from the financial sector, as Wall Street did not trust the technology that rarely delivered on its promises.

Due to this suspicion, there was a significant lack of progress in AI initially. There remained a large gap between the expectations and reality of AI capabilities. Expert systems showed impressive potential when attempting problems that can be seamlessly formalized [10]. For example, in 1997, Deep Blue, IBM's chess-playing program, successfully beat Gary Kasparov, the world chess champion, utilizing a tree-search method to evaluate over 200 million potential moves per second [6]. However, this program could not be successfully applied to a problem that is not as quickly standardized, such as face recognition. For a program to accomplish a task like this, the system must correctly interpret data, learn from it, and apply it to various tasks and goals with flexible adaptation [6].

The need for complex decision-making caused an uproar in AI research. While a few machine learning models are labeled interpretable by design - examples include decision trees, rules, and tables- most AI models function as black boxes, meaning the systems do not reveal sufficient details regarding their internal behavior. [5]. The nature of these opaque decision models will be further discussed in the following section. As AI increasingly intertwines with more human-centric applications, the focus has shifted from accuracy to explainability [11].

In the nascent AI development stages, the primary focus was predominantly on enhancing the accuracy and efficiency of AI models:

- Performance Metrics: Early AI research prioritized performance metrics such as precision and recall, with less consideration for how decisions were made within the model.
- Technological Limitations: Limited by the technology of their times, early developers often had to choose between complex, opaque models that offered better performance and simpler, interpretable ones that did not scale.

While this approach was justified in the early days of AI, when the goal was to establish viable, functional AI systems, today's landscape demands a different paradigm [12]. As AI systems increasingly interact with societal and individual decisions, transparency becomes as critical as accuracy. This shift necessitates a robust exploration of XAI, where understanding and clarifying AI processes are not just an academic interest but a societal imperative [13]. The upcoming sections of this paper will delve into the methodologies and impacts of XAI, seeking to bridge the gap between AI capabilities and human-centric values.

3. What Is XAI?

The field of XAI refers to a wide variety of algorithms. These varying algorithms can be grouped by complexity into three main groups: white, gray, and black box models [14]. White-box models are considered systems with full transparency that do not require extra explainability techniques, such as linear regression [14]. Systems that achieve a more advanced performance but lack interpretability, such as neural networks and random forests, are considered black-box models with high accuracy yet lack transparency [11, 14]. These black boxes are considered opaque models, concealing the methods and algorithms mapping inputs to outputs [15]. For example, an opaque system could emerge when an organization licenses closed-source AI to protect its intellectual property and proprietary AI [15]. The "how" and "why" of the system's process are omitted from the output. Finally, gray-box models fall in between, as they are not intrinsically explainable but can be interpretable when explanation techniques are applied [14].

According to the National Institute of Standards and Technology [16], for a system to be considered explainable, it must possess four fundamental properties:

- Explanation: A system must provide accompanying support or evidence with each decision output.
- Meaningful: The system's explanations are understandable to its intended user, considering different user groups' varying knowledge levels and needs.
- Explanation Accuracy: The system's explanation correctly reflects system processes.
- Knowledge Limits: A system only functions within the range of scenarios and conditions it has been trained for. The system can recognize cases that fall outside its scope.

Knowing what "explainability" means is crucial to understanding the importance of explainable AI. The term does not possess an official definition, but experts have culminated several ways to view the concept of explainability. Explainability describes the type of information provided to users through the user interface to allow informed use of a system's output or recommendation [17]. Explainability answers the simple question, "Why did it do that?".

3.1. Explainability, Interpretability, and Transparency

In many cases, explainability and interpretability are used synonymously; however, according to literature on the topic, interpretability and explainability differ slightly. According to Johnson (2020) and Angelov (2021), the definitions of the terms are as follows:

- Explainability: Relates to the concept of explanation as an interface between AI and humans, including AI systems that are comprehensive to humans through explanation [11].
- Interpretability: The ability to determine cause and effect from a machine learning model that is intrinsically understandable to humans [11, 18].

There are notable qualities that explainable and interpretable systems do and do not possess. The terms used are defined as such:

1. Transparency: The quality of AI systems being understandable by themselves, allowing users to comprehend how the system works [11, 19, 20].
2. User Understanding: the ability of human users to immediately make sense of a system's reasoning and behavior without extra explanations or clarifications [21].
3. Comprehensibility: refers to the capacity of a system or a system's explanations to aid a user in task completion [21].
4. Fairness: The goal that explanations should be egalitarian [21].

Systems can be explainable without being interpretable. Explainability considers explanations of the interface between users and an AI system [11]. Explainability is found in AI systems that are accurate and understandable to humans [11]. In addition, explainability works to clarify its internal decision process to users. It emphasizes the ability of parameters, often hidden in deep neural networks, to justify the results [18, 22]. On the other hand, interpretability relates to how accurately a system can link each cause to an effect [18]. Interpretability describes the capacity of a system to give interpretations in formats understandable to humans. Interpretability also includes to what degree users can understand explanations [23]. For example, deep learning models, such as neural networks, tend to perform highly but lack interpretability [14, 24].

In both interpretable and explainable AI systems, fairness is not guaranteed. Although these techniques provide insight into model behavior and reveal biases, achieving fairness requires the consideration of factors including data bias, algorithmic fairness, and ethical considerations [20]. A system's explainability can be determined by several factors, including complexity, transparency, trust, fidelity, accuracy, and comprehensibility [5, 16, 23]. These dimensions of explainability distinguish explainable systems from black-box models and are critical pieces of explainable AI.

One necessary element of explainable AI is transparency. While explainability answers the question "Why did it do that?" transparency addresses "How does it work?" [25]. In summary, transparency is found in systems that have the potential to be understandable by themselves, making transparent systems the opposite of black box models [11]. Transparency helps lift the lid of black box models. This can reveal a model's structural attributes, evaluation metrics, and descriptive properties from training data to users to foster an understanding of a system's underlying logic [5, 25]. Many machine learning models lacked transparency due to a trade-off between explainability and performance [19]. As previous studies focused on performance improvement, transparency was ignored and placed on the back burner.

AI systems' nontransparent nature began to affect human trust and confidence negatively. More specialized knowledge became necessary to understand AI approaches as the complexity increased. Ordinary users with low algorithmic knowledge found it hard to trust AI systems making crucial decisions, and the lack of transparency hindered user understanding of the exact steps of algorithms [26]. This significantly worsened the problem, as user comprehension of why a specific recommendation is made and how their input affects the results is critical to user satisfaction and trust [26]. For example, in a news recommender system, fair and personalized recommendations give users confidence, leading to trust and continued use [26]. Visible transparency improves search performance, as using explanations improves users' overall satisfaction [26]. In recommender systems, personalization has become a determinant of satisfaction and trust [26]. Moreover, the recommendation explanation sets a prerequisite for a relationship of trust between humans and AI [2]. A lack of transparency in medical applications has been identified as a barrier to AI implementation [23]. Trust in medical AI systems is vital, as the recommendations significantly impact patients' health and well-being [23].

The need for transparency has led to a significant interest in XAI. This field ensures that AI benefits rather than harms society by introducing accountability [3]. Systems that lack transparency don't possess this accountability. In some cases, this is not an issue. For instance, in the historic Go game between Lee Sodel, a highly skilled Go player, and AlphaGo, a DeepMind AI system, AlphaGo made an extremely unexpected move [2]. Experts were unsure why the system made this gaming-altering move. In this case, the nontransparent nature of AlphaGo did not matter, as the

application did not drastically affect human well-being. However, in many applications, the opposite is true.

On the other hand, IBM Watson, a supercomputer containing AI and other analytical software, beat the top players at the game show Jeopardy. This software was then marketed to medical facilities as a cancer-detecting system [2]. When providing results, Watson could not display the reasoning for its output, so patients and doctors could not trust the system [2, 3]. IBM Watson's lack of transparency hindered human trust and was not seen as a successful application. The same mindset can be applied to self-driving cars as well. These automated systems did not react efficiently in a new or unfamiliar environment. In 2018, a computerized vehicle owned by Uber crashed, and the operator was charged with negligent homicide [11]. Transparent and explainable systems are necessary, from a public trust perspective and a legal viewpoint, to provide more reliable and safe systems [11].

The capacity of AI-based systems to elucidate their internal decision-making processes is an area ripe for exploration and innovation:

- Model Transparency: Techniques such as model visualization and feature importance metrics provide insights into the working of complex models, enhancing their transparency.
- Decision Justification: Implementing methods that allow AI systems to justify their decisions can facilitate greater understanding and trust among users.

As AI technologies continue to permeate various aspects of personal and professional life, the ability of these systems to offer clear, understandable explanations for their actions becomes crucial. This supports the development of more robust and reliable AI and upholds the user's right to demand transparency [27]. The next section of this paper will discuss methodologies for formulating these explanations, ensuring that AI systems are effective, accountable, and accessible to the users they serve.

4. Explanations

According to Confalonieri et al., explanations can be understood in two ways: as a line of reasoning or as a problem-solving activity [5]. Viewing explanations as a line of reasoning essentially creates understanding by following the path inference rules take to come to a particular decision [5]. The main issue with this approach was the complexity of explanations, as not all users possess the same knowledge to understand the full extent of explanations thoroughly. This idea was re-conceptualized to approach explanation in a different light: explanations as a problem-solving activity. This altered view not only reconstructs the system's reasoning but also considers various degrees of abstraction, meaning different knowledge levels were considered [28].

Post-hoc and model-based explanations are the most prevalent types when categorizing the explanations provided by XAI systems. Post-hoc methods are commonly used on systems that are not intrinsically interpretable to boost their interpretability [29]. Post-hoc methods do not directly reveal the internal workings of a model. Still, they seek to explain behavior to users by studying outputs and factors that contribute to the result [16]. In other words, explanations are derived after a model makes the prediction. The system uses the nature and attributes of results to generate explanations [17].

On the other hand, model-based explanations focus on the mechanical aspect of recommendations and aim to illustrate how an algorithm suggests a distinct output [5]. Model-based explanation strategies use a different model to explain how the task model functions. The levels of soundness and fidelity are particularly essential for assessing model-based explanations [23]. Model-based explanations are strictly based on the system's underlying assumptions and structure [5]. The following subsections briefly overview post-hoc explanations, addressing different techniques and applications. In addition, several other relevant explanation types, such as self-interpretable models, are referenced.

4.1. Post-hoc Explanations

Post-hoc explainability can be applied in two ways: model-specific and model-agnostic approaches. Model-specific methods produce explanations by utilizing the particular system's internal learning process [30]. Since model-specific interpretability is tailored to bring transparency to specific models, the application will not be suitable for other model types [11, 20, 30]. In contrast, model-agnostic methods are independent of the applied system. Model-agnostic methods develop end-user explanations using the inputs and predictions of the model [20, 30]. The lack of specificity of model-agnostic methods allows for wide-scale usage. In addition, the interpretability of post-hoc models can be further divided into local and global methods.

4.1.1. Local Methods

Local methods obtain explainability by segmenting the solution space and providing less intricate explanations that apply to the entire model [29]. A per-decision or single-decision explanation is the most dominant type of local explanation [16]. It provides insight into the aspects that impact the algorithm's decision for a particular input. Local explanations allow for a local approximation of how a black-box model functions [11]. The most well-known example of local methods is LIME (Local Interpretable Model Agnostic Explainer) [16, 17]. LIME functions by taking a decision and creating an interpretable model that illustrates the local decision, which is then used to deliver per-feature explanations [16]. LIME perturbs training data into a new dataset to form a new interpretable model [11]. Another example of local explanations is SHAP (Shapely Additive exPlanations), which uses a mechanism of additive feature attributions to reveal the significance of input factors [14, 17].

4.1.2. Global Methods

Global methods employ interpretable mechanisms, such as decision trees, to extract a simplified version of a complex black box model to supply understandable explanations for each decision made by the model [11]. This makes it possible to comprehend the behavior of the black-box model and how it relates to its trained characteristics [11]. Global explanations can construct post-hoc explanations

on the whole algorithm [16]. Partial Dependence Plots (PDPs) and Testing with Concept Activation Vectors (TCAV) are examples of global explanations. PDPs demonstrate the modification of predicted responses about altered data components. At the same time, TCAVs explain deep neural networks in a more user-friendly manner and have been applied to image classification systems [16].In addition, a global variant of LIME exists, SP-LIME, which uses applicable local LIME outputs as synopsis explanations [16].

4.2. Self-Interpretable Models

Self-interpretable models are intrinsically explainable, meaning humans can directly understand them. The models are the explanation due to a transparent reasoning process [16, 31, 32]. However, many sources claim self-interpretable models are less accurate than post-hoc explanations due to a trade-off between accuracy and interpretability [16, 33]. The most common self-interpretable models include regression models and decision trees [16, 34].

4.3. Other Explanation Models

In addition, several other explanations exist that do not perfectly fit into a category. The most relevant of these explanation models are defined below.

Forms of Model Explanations:

- Introspective Methods: Explanations are formed by connecting inputs to outputs in black-box models. For example, reflective methods can be applied to image classifications with Deep Neural Networks [5, 35, 36] and [37].
- Counterfactual Methods: Explanations provide "what-if" statements regarding how the outputs of a predicted model could be affected by input changes [5, 38, 39, 40] and [41].
- Explanation by Feature Relevance: A method of post-hoc explainability clarifies a model's internal functioning by calculating a relevance score for each variable. The comparison of scores depicts the weight each variable holds [20, 42] and [43].
- Explanation by Simplification: Explanations that use a trained model to formulate a simplified representation to assemble an easily implementable model. These models optimize similarity to the original model while simultaneously decreasing complexity [11, 29] and [44].

AI-based systems must explain their decisions, which may soon transition from a best practice to a mandatory requirement. This shift is driven both by evolving regulatory frameworks aimed at safeguarding consumer rights and by ethical standards that promote transparency and accountability [45]:

- Regulatory Compliance: Legislations such as the EU's General Data Protection Regulation (GDPR) already impose obligations on AI to explain decisions that affect individuals, signaling a broader trend towards legal mandates.
- Ethical Accountability: Beyond compliance, there is a growing recognition of the ethical obligation for AI to be transparent, particularly in systems that impact public welfare and individual freedoms.

This development is poised to significantly benefit numerous business sectors by enhancing consumer trust, facilitating more informed decision-making, and improving the overall user experience with AI technologies.

5. From ANNs (sub-symbolic) to Rules (symbolic)

Extracting rules from ANNs is crucial in demystifying these models' "black-box" nature, making their decisions understandable and interpretable to humans. This process involves translating the intricate, non-linear relationships learned by the network into a set of rules that humans can easily understand. To illustrate this process, we'll explore a detailed example of how rules can be extracted from an ANN trained on a simplified dataset for predicting loan approval based on applicant features.

5.1. Background

Let us use the example of a fictional financial institution that has created an ANN to evaluate loan applications. The ANN considers various applicant features such as Age, Income, Credit Score, and Employment Status and provides a binary decision: Approve or Deny. Despite the ANN's high accuracy, the decision-making process is not transparent. This makes it challenging for loan officers to explain decisions to applicants or to ensure compliance with regulations. The institution aims to derive understandable rules from the ANN to address this.

5.2. ANN Architecture

The ANN in this example is a simple feedforward network with one hidden layer. The input layer has four neurons corresponding to the applicant features. The hidden layer has a few neurons (say five for simplicity) using ReLU (Rectified Linear Unit) as the activation function [41]. The output layer has one neuron and uses a sigmoid activation function to output a probability of loan approval.

5.3. Rule Extraction Process

The rule extraction process involves several steps designed to translate the ANN's learned weights and biases into a set of if-then rules that replicate the network's decision-making process as closely as possible:

- Simplification: The first step involves simplifying the ANN to make the rule extraction more manageable. This could include pruning insignificant weights (shallows values) and neurons that have little impact on the output based on sensitivity analysis.
- Discretization: Since ANNs deal with continuous inputs and hidden layer activations, a discretization process is applied to convert these continuous values into categorical ranges. For

instance, age might be categorized into 'Young', 'Middle-aged', and 'Old'; Income into 'Low', 'Medium', and 'High'; Credit Score into 'Poor', 'Fair', 'Good', and 'Excellent'; and Employment Status into 'Unemployed' and 'Employed'.

- Activation Pattern Analysis: Next, the activation patterns of the neurons in the hidden layer are analyzed for each input pattern. This involves feeding various combinations of the discretized input variables into the simplified network and observing which neurons in the hidden layer are activated for each combination. An activation threshold is defined to determine whether a neuron is considered activated.

- Rule Generation: Based on the activation patterns observed, rules are generated to replicate the ANN's decision process. Each rule corresponds to a path from the input layer through the activated hidden neurons to the output decision. For example:

  - If (Age is Young) and (Income is High) and (Credit Score is Good) and (Employment Status is Employed), then Approve Loan.
  - If (Age is Middle-aged) and (Credit Score is Poor), then Deny Loan.

  This step involves identifying which combinations of input features and hidden neuron activations lead to loan approval or denial, effectively translating the ANN's complex decision boundaries into more interpretable formats.

- Rule Refinement and Validation: The initial set of rules may be too complex or too numerous for practical use. Rule refinement techniques simplify and consolidate the rules without significantly reducing their accuracy in replicating the ANN's decisions. The refined rules are then validated against a test dataset to accurately reflect the ANN's behavior. This may involve adjusting the rules based on misclassifications or applying techniques to handle exceptions and edge cases.

After applying the rule extraction process to our hypothetical ANN, we might end up with a set of simplified, human-readable rules such as:

- Rule 1: If (Income is High) and (Credit Score is Excellent), then Approve Loan.
- Rule 2: If (Employment Status is Unemployed) and (Credit Score is Poor or Fair), then Deny Loan.
- Rule 3: If (Age is Old) and (Income is Low) and (Employment Status is Employed), then Deny Loan.

These rules provide clear criteria derived from the ANN's learned patterns, making the decision-making process transparent and justifiable.

5.4. Advantages and Challenges

Some advantages include:

- Transparency: The extracted rules make the ANN's decisions transparent and understandable to humans.
- Compliance: Clear rules can help ensure compliance with regulatory requirements for explainable AI.
- Trust: Understanding how decisions are made can increase user trust in the AI system.

Some challenges are:

- Complexity: The rule extraction process can be complex, especially for deep or highly non-linear networks [46].
- Approximation: The extracted rules approximate the ANN's decision process and may not capture all nuances.
- Scalability: Extracting rules from large, deep neural networks with many inputs and hidden layers can be challenging and may result in many complex rules [47].

5.5. Summary

Extracting rules from ANNs makes AI decision-making transparent, understandable, and justifiable. Although there are challenges, especially with complex networks, this process is crucial for responsible and ethical AI use. By making AI systems more interpretable, we can establish trust with users, ensure compliance with regulations, and gain valuable insights into decision-making.

6. Fuzzy Cognitive Maps

The pendulum in AI is swinging back from purely statistical approaches toward integrating structured knowledge. FCMs are powerful cognitive tools for modeling and simulating complex systems. They blend elements from artificial neural networks, graph theory, and semantic nets to offer a unique approach to understanding and predicting system behavior. FCMs incorporate the concept of fuzziness from fuzzy logic, enabling them to handle ambiguity and uncertainty inherent in real-world scenarios. This extensive report delves into the origins of FCMs, provides illustrative case studies, and discusses their advantages and disadvantages, with references to their similarities to artificial neural networks, graphs, and semantic nets [48].

6.1. Origins

Bart Kosko introduced the concept of FCMs in the 1980s as an extension of cognitive maps. Cognitive maps, developed by Axelrod, were diagrams that represented beliefs and their interconnections. Kosko's introduction of fuzziness to these maps allowed for the representation of causal reasoning with degrees of truth rather than binary true/false values, thus capturing the uncertain and imprecise nature of human knowledge and decision-making processes. FCMs combine elements from fuzzy logic, introduced by Lotfi A. Zadeh, with the structure of cognitive maps to model complex systems.

6.2. Structure and Functionality

FCMs are graph-based representations where nodes represent concepts or entities within a system, and directed edges depict the causal relationships between these concepts. Each edge is assigned a weight that indicates the relationship's strength and direction (positive or negative). This structure closely mirrors that of artificial neural networks, particularly in how information flows through the network and how activation levels of concepts are updated based on the input they receive, akin to the weighted connections between neurons in neural networks [49].

However, unlike typical neural networks that learn from data through backpropagation or other learning algorithms, the weights in FCMs are often determined by experts or derived from data using specific algorithms designed for FCMs. The concepts in FCMs can be activated like neurons, with their states updated based on fuzzy causal relations, allowing for dynamic modeling of system behavior over time. Integrating structured knowledge graphs with distributed neural network representations offers a promising path to augmented intelligence. We get the flexible statistical power of neural networks that predict, classify, and generate based on patterns—combined with the formalized curated knowledge encoding facts, logic, and semantics via knowledge graphs [50].

6.3. The Inherent Reasoning Mechanism

The primary function of the reasoning rule in FCM models is to update the activation values of concepts iteratively, starting from initial conditions and continuing until a stopping criterion is satisfied. During each iteration, the reasoning rule utilizes three primary components to conduct these calculations: the weight matrix, which signifies the connections between concepts; the activation values of concepts from the previous iteration; and the activation function.

Eq. (1) shows a general rule commonly found in FCMs-related papers:

$$a_i^{(t)} = f\left(\sum_{j=1, i\neq j}^{N} a_j^{(t-1)} w_{ji}\right), \tag{1}$$

Recently, in [51], the author proposed an updated quasi-nonlinear reasoning rule depicted in Eq. (2):

$$a_i^{(t)} = \underbrace{\phi \cdot f\left(\sum_{j=1}^{N} a_j^{(t-1)} w_{ji}\right)}_{\text{nonlinear component}} + \underbrace{(1-\phi)\cdot a_i^{(0)}}_{\text{linear component}}, \tag{2}$$

such that $0 \leq \phi \leq 1$ is the nonlinearity coefficient. When $\phi = 1$, the concept's activation value depends on the activation values of connected concepts in the previous iteration. When $0 < \phi < 1$, we add a linear component to the reasoning rule devoted to preserving the initial activation values of concepts. When $\phi = 0$, the model narrows down to a linear regression where the initial activation values of concepts act as regressors. In their paper, Nápoles et al. [51] used the quasi-nonlinear reasoning rule to quantify implicit bias in pattern classification datasets. In contrast, the authors in [41] resorted to this rule to develop a recurrence-aware FCM-based classifier.

6.4. How Activation Functions Work

The activation function $f : \mathbb{R} \rightarrow I$ is an essential component in the reasoning rule of FCM-based models. This monotonically non-decreasing function keeps the activation value of each concept within the desired image set I, which can be discrete (a finite set) or continuous (a numeric-valued interval). It should be mentioned that I must be bounded; otherwise, the reasoning rule could explode due to the successive additions and multiplications when updating concepts' activation values during reasoning. Table **??** portrays relevant activation functions found in the literature.

6.5. Relevant Case Studies

For illustration purposes, Figure 1 shows an example of an FCM created to model a case of autism [32]. FCMs have been applied across various domains, demonstrating their versatility and effectiveness as a hybrid AI tool:

- Decision Support Systems: FCMs model complex decision-making processes, integrating expert knowledge and data-driven insights to support decisions in healthcare, environmental management, and business strategy.

- Predictive Modeling: In healthcare, FCMs model the progression of diseases or the impact of treatments, incorporating medical expertise and patient data to predict outcomes and support personalized medicine [52].

- System Analysis and Design: FCMs help analyze and design complex systems, such as socio-economic systems or ecosystems, by modeling the interactions between various factors and predicting the impact of changes or interventions.

- Healthcare Management: FCMs have been employed to model and predict patient outcomes in healthcare settings. For example, an FCM can be developed to understand the complex interplay between patient symptoms, treatment options, and possible outcomes, aiding medical professionals in decision-making [53].

- Environmental and Ecological Systems: In environmental studies, FCMs have been used to model the impact of human activities on ecosystems, allowing for the simulation of various scenarios based on different policies or interventions. This application showcases the strength of FCMs in handling systems where data may be scarce or imprecise [54].

- Business and Strategic Planning: FCMs assist in strategic planning and decision-making within business contexts by modeling the relationships between market forces, company policies, and financial outcomes, offering a tool for scenario analysis and strategy development [55].

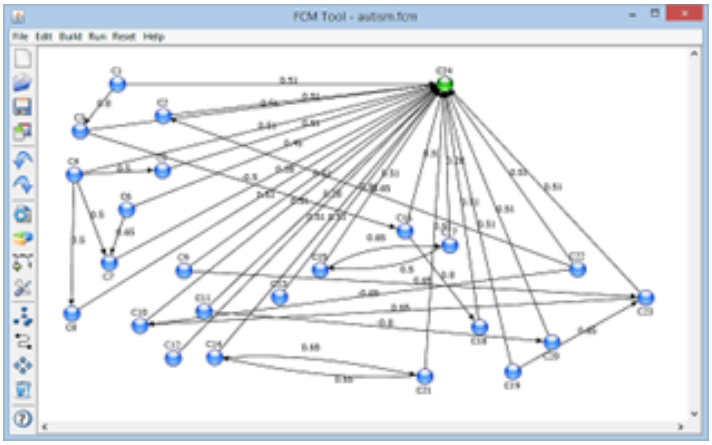

Figure 1: Real example created with FCM Tool.

6.6. Advantages

The hybrid nature of FCMs offers several advantages:

- Interpretability and Transparency: The symbolic representation of concepts and causal relationships in FCMs provides clarity and understandability, facilitating communication with experts and stakeholders and supporting explainable AI.
- Flexibility and Adaptability: FCMs can be easily updated with new knowledge or data, allowing them to adapt to changing conditions or insights. This makes them particularly valuable in fields where knowledge evolves rapidly.
- Handling of Uncertainty: Using fuzzy values to represent causal strengths enables FCMs to deal effectively with uncertainty and ambiguity, providing more nuanced and realistic modeling of complex systems [4].
- Integration of Expert Knowledge and Data-Driven Insights: FCMs uniquely combine expert domain knowledge with learning from data, bridging the gap between purely knowledge-driven and purely data-driven approaches.
- Interpretability: The graphical representation of FCMs, similar to semantic nets, allows for straightforward interpretation and understanding of the modeled system, making it accessible to experts and stakeholders without deep technical knowledge of AI.
- Flexibility: FCMs can incorporate quantitative and qualitative data, effectively handling uncertainty and imprecision through fuzzy logic. This flexibility makes them suitable for a wide range of applications.
- Dynamic Modeling Capability: FCMs can simulate the dynamic behavior of systems over time, providing valuable insights into potential future states based on different inputs or changes in the system [56].

6.7. Limitations

Despite their advantages, FCMs also face several challenges:

- Complexity with Large Maps: As the number of concepts and relationships in an FCM increases, the map can become complex and challenging to manage, analyze, and interpret [57].
- Learning and Optimization: While FCMs can learn from data, adjusting the fuzzy values of causal relationships can be computationally intensive and require sophisticated optimization techniques, especially for large and complex maps [58].
- Quantification of Expert Knowledge: Translating expert knowledge into precise fuzzy values for causal relationships can be challenging and may introduce subjectivity, requiring careful validation and sensitivity analysis [59].
- Subjectivity in Model Construction: The reliance on expert knowledge for constructing FCMs can introduce subjectivity, especially in determining the strength and direction of causal relationships between concepts.
- Complexity with Large Maps: As the number of concepts increases, the FCM can become complex and challenging to manage and interpret, potentially requiring sophisticated computational tools for simulation and analysis.
- Limited Learning Capability: While FCMs can be adjusted or trained based on data to some extent, they lack the deep learning capabilities of more advanced neural networks, which can autonomously learn complex patterns from large datasets [60].

7. Applications

Numerous potential applications exist for XAI techniques and models, including healthcare, law, data science, and business [55]. This section explores the need for explainability in these applications, including their current uses, limitations, and future development.

7.1. Healthcare

In healthcare, there are many applications of XAI such as diagnosis, treatment recommendations, and surgery [23, 61, 62]. For example, an explainable model was proposed for diagnosing skin diseases. Using saliency maps to highlight important parts of the image crucial to diagnosis, dermatologists can easily understand the model's arrival at a diagnosis and then provide a more in-detail diagnosis [61]. According to a survey by Zhang et al., LIME is the most commonly used XAI approach in medical applications [62].

Throughout the COVID-19 pandemic, AI has shown potential in developing solutions to confront the difficulties presented by the virus [61]. However, the lack of transparency in black-box models has hindered their acceptance in clinical practice. With the development in user trust and model performance, XAI can attempt these problems in the future [61]. XAI techniques have been created in the context of medical image analysis to facilitate disease detection and diagnosis through feature visualization [61]. This allows medical professionals and their patients to obtain a deeper insight into the model's process, building confidence in its accuracy. In high-stakes applications, specifically healthcare, there is debate about whether

explainable modeling is necessary. To some, explainability is crucial. On the other hand, some say prioritizing explainability above accuracy in healthcare systems can be unethical [23]. According to [23], the post-hoc explanations can be delusive, but a potential solution is to create post-hoc explanation models with argumentative support.

Suppose the case of an ANN equipped with a rule extraction method can be deployed to diagnose diseases from medical imaging with high accuracy. The ANN processes complex imaging data to identify patterns indicative of specific conditions, such as tumors in MRI scans. A rule extraction technique is integrated into the system to ensure clinicians and patients understand the diagnostic process. This technique translates the ANN's intricate decision-making into simple, interpretable rules, such as the presence of specific shapes or textures associated with malignancy. This not only aids medical professionals in making informed treatment decisions but also enhances patient trust by providing clear explanations for the diagnoses made by the AI system.

7.2. Law

In the context of legal applications, XAI possesses several potential applications. As stated by Reddy et al., XAI can be used for legal document analysis, contract review, legal decision-making, and addressing challenges in legal domains [61]. AI can help analyze large volumes of legal documents and sort significant information to facilitate a more accurate analysis, as well as assist in recommending plea bargains or predicting case outcomes [61, 63]. Despite the increasing emphasis on AI in the legal world, systems still struggle to perform at necessary levels due to the precise nature of legal work. Such characteristics include the exact nature of legal jargon, the high level of expertise required, the mass amount of situational exceptions, and the limited tolerance of mistakes [61]. The motivation for interpretable, explainable, and trustworthy systems feeds the recent upsurge of XAI research in legal applications.

In legal applications, an FCM can be a sophisticated tool for modeling and visualizing the intricate dynamics of legal cases and legislative processes. By capturing and representing the causal relationships between various legal factors—such as statutes, precedents, and evidentiary variables—FCMs enable legal professionals to simulate and scrutinize the potential outcomes of different legal strategies in a visually interpretable format. This capability goes beyond basic explainability by showing outcomes and allowing users to interact with the map to adjust variables and immediately see different scenario outcomes. This interactive, interpretable visualization aids in understanding complex legal interdependencies, facilitating more informed decision-making and strategy formulation, especially in cases involving overlapping laws and diverse outcomes.

7.3. Finance

In the financial sector, the applications of XAI can be split into thematic categories. These clusters include financial distress and corporate failure, algorithmic and high-frequency trading, forecasting/predictive analysis, text mining and sentiment analysis, financial fraud, pricing and valuation, scheduling, and investor behavior [64]. In addition, Reddy et al. describe the potential applications of XAI in finance as follows [61]:

- **Fraud Detection:** Explain decisions by identifying the reasons behind fraudulent activities and prevent future issues.
- **Credit Scoring:** Allows banks and their customers to understand exactly why a particular credit score was calculated and facilitates lending decisions.
- **Investment Management:** Increased transparency in portfolio management can lead to better performance and more satisfied investors.
- **Compliance:** XAI could assist in mitigating potential biases and avoiding legal issues.
- **Customer Service:** XAI will improve customer service by, for example, including explanations along with loan denials to improve customer understanding and satisfaction.

According to additional literature on the topic, subjects within the finance domain commonly discussed as potential applications of XAI include risk management, portfolio optimization, electronic financial transaction clarification, and anti-money laundering [64]. Due to the high level of regulations in financial domains, XAI is necessary to augment processes to ensure trust and transparency and mitigate risks [65].

Suppose the case of an ANN equipped with a rule extraction method can be effectively used for credit scoring. The ANN analyzes extensive data sets, including transaction history, payment behavior, and credit utilization, to assess the creditworthiness of applicants. By integrating a rule extraction method, the system can transparently generate and provide clear, human-understandable rules that explain its credit-scoring decisions. This transparency not only aids financial analysts in understanding the model's decision-making process but also ensures compliance with regulatory requirements regarding fairness and explainability in credit assessments.

An FCM can model and visualize a client's financial stability or market for the same finance application. By representing elements like market trends, economic indicators, and individual financial behaviors as nodes and their interdependencies as edges, FCMs allow financial analysts to simulate and interpret complex financial scenarios. This method provides a dynamic, interpretable visualization beyond mere explanation, enabling interactive exploration of potential financial outcomes based on varying inputs. Such interpretability is invaluable in strategic financial planning and risk assessment, allowing the decision-makers to foresee and mitigate potential financial instabilities or crises.

8. Future

As complex and human-centric systems become more prevalent, there is a growing need for explainable AI in many applications. Due to the rapid increase in AI, there are currently few regulations and rules governing these systems. However, as the need for trust and transparency continues to rise, regulations are essential to ensure both ethical and accountable AI.

8.1. Current Regulations

Historically, AI-based systems have operated in an environment with minimal regulatory oversight regarding their need to explain internal decision-making processes:

- Early AI Developments: Initially, AI technologies were developed and deployed with a focus on functionality and performance, often at the expense of transparency and accountability [66].

- Regulatory Lag: There has been a significant lag in developing and implementing regulations that require AI systems to be explainable, partly due to the rapid pace of technological advancement outstripping policy development.

However, as the implications of AI technologies have become more apparent, there is a growing consensus among government bodies and policymakers about the necessity of regulatory frameworks that ensure AI systems are transparent and accountable. This shift reflects a broader awareness of AI's potential impacts on society and the need for appropriate safeguards.

The regulation of AI is becoming extremely important in terms of ethics and responsible decision-making. The European Union's General Data Protection Regulation (GDPR) was put into effect in 2018, and the GDPR has raised several legal and ethical questions regarding safety, responsibility, malfunction liability, and the overall trade-offs associated with AI decisions [67]. The GDPR gives citizens a "right to explanation" in algorithmic choices that significantly affect them [68, 69]. Regulations like the GDPR make it nearly impossible to use black-box models in various sectors, emphasizing the growing need for explainability and transparency [70, 71]. Additionally, the EU's intense regulatory actions involving digital markets, including the AI domain, strive to provide an ethical approach to AI applications [72]. Additionally, Hacker (2023) highlights the transformative prospects as well as risks associated with large generative AI models (LGAIMs), such as ChatGPT, and how current regulations are not suited to manage this class of AI [73].

In April 2021, the European Commission proposed a groundbreaking proposal for the first-ever EU regulatory framework for AI. This framework consists of a risk-based classification technique in which the level of risk specifies the regulation applied to a system [74]. The AI Act manages the opacity of particular systems, emphasizing systems classified as high-risk through a focus on transparency [75]. If implemented, the AI Act will represent the world's baseline rules for overseeing AI. Furthermore, generative AI systems such as ChatGPT must follow transparency conditions, such as publishing data synopses for training the system [74].

In summary, AI regulations are developing to address ethical considerations, transparency, and the responsible use of AI across diverse sectors. The GDPR and corresponding endeavors emphasize the demand for transparency and accountability in AI decision-making. At the same time, ongoing discussions in the EU seek to shape AI development in a human-centric and ethical fashion.

8.2. The Future of XAI

The future of XAI holds tremendous promise and challenges. In an increasingly AI-driven world, the possible applications are extensive; however, awareness of the fragile nature and potential biases within AI systems is expanding. As stated previously, global organizations are attempting to craft standards for responsible AI to mitigate concerns. These regulations strive to make AI systems exemplify more transparency and accountability, making the demand for explainable systems higher than ever.

As different organizations and governments pass regulations, the dilemma now shifts: Is regulating the AI available to specific users and not others ethical? When tackling this issue, enforcing rules on AI is essential. Without universal regulations, organizations may pass conflicting laws, which could immensely harm companies attempting to operate systems globally. For example, with search engines experimenting with generative AI systems, such as Google's Bard or Gemini, non-universal regulations would require several system versions to adhere to local regulations, causing unnecessary complexities. Moreover, universal regulations would provide businesses with legal certainty. Ethically, universal regulations will form a standard for ethical AI, assisting in eliminating biased and discriminatory systems. This will also allow users to feel more trust in consistently observed systems, leading to increased adoption of systems.

In conclusion, from a business perspective, the universal enforcement of AI regulations provides many advantages. Companies should prioritize accountable AI and support coordinated regulations to develop ethical, transparent, and innovative AI technologies. Explainable systems are the key to the future of Responsible AI.

References

[1] V. K. Michael Chui, B. McCarthy, "An Executives Guide to AI," McKinsey & Company, 2018.

[2] W. Samek, T. Wiegand, K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," arXiv preprint arXiv:1708.08296, 2017, doi:10.48550/arXiv.1708.08296.

[3] "Unlocking the black box with explainable AI - Infosys," Infosys, 2019.

[4] M. Leon, "Aggregating Procedure for Fuzzy Cognitive Maps," The International FLAIRS Conference Proceedings, **36**(1), 2023, doi:10.32473/flairs.36.133082.

[5] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, "A historical perspective of explainable Artificial Intelligence," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **11**(1), e1391, 2021, doi:10.1002/widm.1391.

[6] M. Haenlein, A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," California Management Review, **61**(4), 5–14, 2019, doi:10.1177/0008125619864925.

[7] A. Abraham, "Rule-Based expert systems," Handbook of Measuring System Design, 2005, doi:10.1002/9780470027325.s6405.

[8] T. G. Gill, "Early expert systems: Where are they now?" MIS Quarterly, **19**(1), 51–81, 1995, doi:10.2307/249711.

[9] J. Kastner, S. Hong, "A review of expert systems," European Journal of Operational Research, **18**(3), 285–292, 1984, doi:10.1016/0377-2217(84)90202-0.

[10] E. Struble, M. Leon, E. Skordilis, "Intelligent Prevention of DDoS Attacks using Reinforcement Learning and Smart Contracts," The International FLAIRS Conference Proceedings, **37**(1), 2024, doi:10.32473/flairs.37.1.135349.

[11] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, P. M. Atkinson, "Explainable artificial intelligence: an analytical review," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **11**(5), e1424, 2021, doi:10.1002/widm.1424.

[12] M. Leon, "Fuzzy Cognitive Maps as a Bridge between Symbolic and Subsymbolic Artificial Intelligence," International Journal on Cybernetics & Informatics (IJCI). 3rd International Conference on Artificial Intelligence Advances (AIAD 2024), **13**(4), 57–75, 2024, doi:10.5121/ijci.2024.130406.

[13] M. Leon, L. Mkrtchyan, B. Depaire, D. Ruan, K. Vanhoof, "Learning and clustering of fuzzy cognitive maps for travel behaviour analysis," Knowledge and Information Systems, **39**(2), 435–462, 2013, doi:10.1007/s10115-013-0616-z.

[14] M. Schemmer, N. Kühl, G. Satzger, "Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence," arXiv preprint arXiv:2109.13827, 2021, doi:10.48550/arXiv.2109.13827.

[15] D. Doran, S. Schulz, T. R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," arXiv preprint arXiv:1710.00794, 2017, doi:10.48550/arXiv.1710.00794.

[16] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, M. A. Przybocki, "Four principles of explainable artificial intelligence," Gaithersburg, Maryland, 18, 2020, doi:10.6028/NIST.IR.8312.

[17] P. Bhattacharya, N. Ramesh, "Explainable AI: A Practical Perspective," Infosys, 2020.

[18] J. Johnson, "Interpretability vs explainability: The black box of machine learning," BMC Blogs, 2020.

[19] D. Minh, H. X. Wang, Y. F. Li, T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," Artificial Intelligence Review, **55**, 3503–3568, 2022, doi:10.1007/s10462-021-10088-y.

[20] A. Rai, "Explainable AI: From black box to glass box," Journal of the Academy of Marketing Science, **48**, 137–141, 2020, doi:10.1007/s11747-019-00710-5.

[21] C. Meske, E. Bunde, J. Schneider, M. Gersch, "Explainable Artificial Intelligence: Objectives, Stakeholders and Future Research Opportunities," Information Systems Management, 2020, doi:10.1080/10580530.2020.1849465.

[22] S. S. Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, H.-W. Liang, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," Informatics in Medicine Unlocked, **40**, 101286, 2023, doi:10.1016/j.imu.2023.101286.

[23] A. F. Markus, J. A. Kors, P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," Journal of Biomedical Informatics, **113**, 103655, 2021, doi:10.1016/j.jbi.2020.103655.

[24] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, "Explainable AI: the new 42?" in Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2, 295–303, Springer, 2018, doi:10.1007/978-3-319-99740-7_21.

[25] C. Oxborough, E. Cameron, A. Rao, A. Birchall, A. Townsend, C. Westermann, "Explainable AI: Driving business value through greater understanding," Retrieved from PWC website: https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf, 2018.

[26] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," International Journal of Human-Computer Studies, **146**, 102551, 2021, doi:10.1016/j.ijhcs.2020.102551.

[27] G. Nápoles, M. L. Espinosa, I. Grau, K. Vanhoof, R. Bello, Fuzzy cognitive maps based models for pattern classification: Advances and challenges, volume 360, 83–98, Springer Verlag, 2018.

[28] G. Nápoles, M. Leon, I. Grau, K. Vanhoof, "FCM Expert: Software Tool for Scenario Analysis and Pattern Classification Based on Fuzzy Cognitive Maps," International Journal on Artificial Intelligence Tools, **27**(07), 1860010, 2018, doi:10.1142/S0218213018600102.

[29] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, **58**, 82–115, 2020, doi:10.1016/j.inffus.2019.12.012.

[30] D. Vale, A. El-Sharif, M. Ali, "Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law," AI and Ethics, **2**, 815–826, 2022, doi:10.1007/s43681-022-00142-y.

[31] M. Xue, Q. Huang, H. Zhang, L. Cheng, J. Song, M. Wu, M. Song, "Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition," arXiv preprint arXiv:2208.10431, 2022, doi:10.48550/arXiv.2208.10431.

[32] M. Leon Espinosa, G. Napoles Ruiz, "Modeling and Experimentation Framework for Fuzzy Cognitive Maps," Proceedings of the AAAI Conference on Artificial Intelligence, **30**(1), 2016, doi:10.1609/aaai.v30i1.9841.

[33] A. Adadi, M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," IEEE Access, **6**, 52138–52160, 2018, doi:10.1109/ACCESS.2018.2870052.

[34] V. G. Costa, C. E. Pedreira, "Recent advances in decision trees: An updated survey," Artificial Intelligence Review, **56**(5), 4765–4800, 2023, doi:10.1007/s10462-022-10275-5.

[35] J. F. Allen, S. Schmidt, S. A. Gabriel, "Uncovering Strategies and Commitment Through Machine Learning System Introspection," SN Computer Science, **4**(4), 322, 2023, doi:10.1007/s42979-023-01747-8.

[36] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, "Explainability in deep reinforcement learning," Knowledge-Based Systems, **214**, 106685, 2021, doi:10.48550/arXiv.2008.06693.

[37] P. Sequeira, E. Yeh, M. T. Gervasio, "Interestingness Elements for Explainable Reinforcement Learning through Introspection." in IUI workshops, volume 1, 2019, doi:10.48550/arXiv.1912.09007.

[38] X. Dai, M. T. Keane, L. Shalloo, E. Ruelle, R. M. Byrne, "Counterfactual explanations for prediction and diagnosis in XAI," in Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 215–226, 2022, doi:10.1145/3514094.3534144.

[39] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, "If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques," arXiv preprint arXiv:2103.01035, 2021, doi:10.48550/arXiv.2103.01035.

[40] G. Warren, M. T. Keane, R. M. Byrne, "Features of Explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI," arXiv preprint arXiv:2204.10152, 2022, doi:10.48550/arXiv.2204.10152.

[41] G. Nápoles, Y. Salgueiro, I. Grau, M. Leon, "Recurrence-Aware Long-Term Cognitive Network for Explainable Pattern Classification," IEEE Transactions on Cybernetics, **53**(10), 6083–6094, 2023, doi:10.48550/arXiv.2107.03423.

[42] P. A. Moreno-Sanchez, "An automated feature selection and classification pipeline to improve explainability of clinical prediction models," in 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), 527–534, IEEE, 2021, doi:10.1109/ICHI52183.2021.00100.

[43] J. Tritscher, A. Krause, A. Hotho, "Feature relevance XAI in anomaly detection: Reviewing approaches and challenges," Frontiers in Artificial Intelligence, **6**, 1099521, 2023, doi:10.3389/frai.2023.1099521.

[44] J. Tritscher, M. Ring, D. Schlr, L. Hettinger, A. Hotho, "Evaluation of post-hoc XAI approaches through synthetic tabular data," in Foundations of Intelligent Systems: 25th International Symposium, ISMIS 2020, Graz, Austria, September 23–25, 2020, Proceedings, 422–430, Springer, 2020, doi:10.1007/978-3-030-59491-6_40.

[45] G. Nápoles, F. Hoitsma, A. Knoben, A. Jastrzebska, M. Leon, "Prolog-based agnostic explanation module for structured pattern classification," Information Sciences, **622**, 1196–1227, 2023, doi:10.1016/j.ins.2022.12.012.

[46] Z. Yang, J. Liu, K. Wu, "Learning of Boosting Fuzzy Cognitive Maps Using a Real-coded Genetic Algorithm," in 2019 IEEE Congress on Evolutionary Computation (CEC), 966–973, 2019, doi:10.1109/CEC.2019.8789975.

[47] W. Liang, Y. Zhang, X. Liu, H. Yin, J. Wang, Y. Yang, "Towards improved multifactorial particle swarm optimization learning of fuzzy cognitive maps: A case study on air quality prediction," Applied Soft Computing, **130**, 109708, 2022, doi:10.1016/j.asoc.2022.109708.

[48] Y. Hu, Y. Guo, R. Fu, "A novel wind speed forecasting combined model using variational mode decomposition, sparse auto-encoder and optimized fuzzy cognitive mapping network," Energy, **278**, 127926, 2023, doi:10.1016/j.energy.2023.127926.

[49] W. Hoyos, J. Aguilar, M. Toro, "A clinical decision-support system for dengue based on fuzzy cognitive maps," Health Care Management Science, **25**(4), 666–681, 2022, doi:10.1007/s10729-022-09611-6.

[50] W. Hoyos, J. Aguilar, M. Toro, "PRV-FCM: An extension of fuzzy cognitive maps for prescriptive modeling," Expert Systems with Applications, **231**, 120729, 2023, doi:10.1016/j.eswa.2023.120729.

[51] G. Nápoles, I. Grau, L. Concepción, L. K. Koumeri, J. P. Papa, "Modeling implicit bias with fuzzy cognitive maps," Neurocomputing, **481**, 33–45, 2022.

[52] K. Poczeta, E. I. Papageorgiou, "Energy Use Forecasting with the Use of a Nested Structure Based on Fuzzy Cognitive Maps and Artificial Neural Networks," Energies, **15**(20), 7542, 2022, doi:10.3390/en15207542.

[53] G. D. Karatzinis, N. A. Apostolikas, Y. S. Boutalis, G. A. Papakostas, "Fuzzy Cognitive Networks in Diverse Applications Using Hybrid Representative Structures," International Journal of Fuzzy Systems, **25**(7), 2534–2554, 2023, doi:10.1007/s40815-023-01564-4.

[54] O. Orang, P. C. de Lima e Silva, F. G. Guimarães, "Time series forecasting using fuzzy cognitive maps: a survey," Artificial Intelligence Review, **56**, 7733–7794, 2023, doi:10.1007/s10462-022-10319-w.

[55] M. Leon, "Business Technology and Innovation Through Problem-Based Learning," in Canada International Conference on Education (CICE-2023) and World Congress on Education (WCE-2023), Infonomics Society, 2023, doi:10.20533/cice.2023.0034.

[56] E. Jiya, O. Georgina, A. O., "A Review of Fuzzy Cognitive Maps Extensions and Learning," Journal of Information Systems and Informatics, **5**(1), 300–323, 2023, doi:10.51519/journalisi.v5i1.447.

[57] R. Schuerkamp, P. J. Giabbanelli, "Extensions of Fuzzy Cognitive Maps: A Systematic Review," ACM Comput. Surv., **56**(2), 53:1–53:36, 2023, doi:10.1145/3610771.

[58] S. Yang, J. Liu, "Time-Series Forecasting Based on High-Order Fuzzy Cognitive Maps and Wavelet Transform," IEEE Transactions on Fuzzy Systems, **26**(6), 3391–3402, 2018, doi:10.1109/TFUZZ.2018.2831640.

[59] T. Koutsellis, G. Xexakis, K. Koasidis, N. Frilingou, A. Karamaneas, A. Nikas, H. Doukas, "In-Cognitive: A web-based Python application for fuzzy cognitive map design, simulation, and uncertainty analysis based on the Monte Carlo method," SoftwareX, **23**, 2023, doi:10.1016/j.softx.2023.101513.

[60] D. Qin, Z. Peng, L. Wu, "Deep attention fuzzy cognitive maps for interpretable multivariate time series prediction," Knowledge-Based Systems, **275**, 110700, 2023, doi:10.1016/j.knosys.2023.110700.

[61] G. P. Reddy, Y. P. Kumar, "Explainable AI (XAI): Explained," in 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 1–6, IEEE, 2023, doi:10.1109/eStream.2023.00001.

[62] Y. Zhang, Y. Weng, J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," Diagnostics, **12**(2), 237, 2022, doi:10.3390/diagnostics12020237.

[63] A. Nielsen, S. Skylaki, M. Norkute, A. Stremitzer, "Effects of XAI on Legal Process," ICAIL '23: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, 2023, doi:10.1145/3593013.3594067.

[64] P. Weber, K. V. Carl, O. Hinz, "Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature," Management Review Quarterly, **74**, 867–907, 2023, doi:10.1007/s11301-023-00320-0.

[65] H. DeSimone, M. Leon, "Explainable AI: The Quest for Transparency in Business and Beyond," in 2024 7th International Conference on Information and Computer Technologies (ICICT), 1–6, IEEE, 2024, doi:10.1109/icict62343.2024.00093.

[66] G. Nápoles, J. L. Salmeron, W. Froelich, R. Falcon, M. Leon, F. Vanhoenshoven, R. Bello, K. Vanhoof, Fuzzy Cognitive Modeling: Theoretical and Practical Considerations, 77–87, Springer Singapore, 2019, doi:10.1007/978-981-13-8311-3_7.

[67] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in International Cross-Domain Conference for Machine Learning and Knowledge Extraction, 1–16, Springer, 2020, doi:10.1007/978-3-030-57321-8_1.

[68] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II, 563–574, Springer, 2019, doi:10.1007/978-3-030-32236-6_51.

[69] W. Saeed, C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," Knowledge-Based Systems, **263**, 110273, 2023, doi:10.48550/arXiv.2111.06420.

[70] A. Toosi, A. G. Bottino, B. Saboury, E. Siegel, A. Rahmim, "A brief history of AI: how to prevent another winter (a critical review)," PET Clinics, **16**(4), 449–469, 2021, doi:10.1016/j.cpet.2021.07.001.

[71] T. Hulsen, "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare," AI, **4**(3), 652–666, 2023, doi:10.3390/ai4030034.

[72] R. Justo-Hanani, "The politics of Artificial Intelligence regulation and governance reform in the European Union," Policy Sciences, **55**(1), 137–159, 2022, doi:10.1007/s11077-022-09452-8.

[73] P. Hacker, A. Engel, M. Mauer, "Regulating ChatGPT and other large generative AI models," in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1112–1123, 2023, doi:10.1145/3593013.3594067.

[74] E. Parliament, "EU AI Act: first regulation on artificial intelligence," 2023.

[75] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, et al., "The role of explainable AI in the context of the AI Act," in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1139–1150, 2023, doi:10.1145/3593013.3594069.