Deep Science Publishing
https://doi.org/10.70593/978-93-49910-74-4_10

DeepScience
Open Access Books

Chapter 10: explainable artificial Advisory and consulting solutions powered by intelligence

10.1. Introduction to Explainable AI

The rapid integration of artificial intelligence (AI) assistance in a plethora of business and lifestyle domains presents a pressing demand for increasing AI transparency and accountability to ensure user trust, regulatory support, and system design integrity. Explainable AI (XAI) is an emerging research domain that traverses the intersection of computer and social sciences, providing innovative methods focused on increasing transparency and interpretability of AI systems, particularly concerning biased algorithmic action and decision production that affects user well-being. Despite the technological advancements present in the unprecedented accuracy improvements of AI systems, limitations remain regarding the transparency of AI functionalities and decision processes, and the inherent biases riddling those aspects. There is significant public interest, concern, and belief regarding AI transparency; however, there exists a restrained cross-section of users who believe they would benefit from AI systems and action decisions being fully explainable and interpretable (Caruana et al., 2000; Lundberg & Lee, 2017; Chen & Zhao, 2020).

Current systems deploy a modicum of transparent design solutions that assume transparency is a one-size-fits-all feature. Without exception, these methods are superficial and nascent. Interestingly enough, there is an observable information asymmetry that exists between governments, policy-makers, and businesses that seek to adopt permissive regulations around the use of AI. One of the common approaches being utilized to address the innate concerns of these parties is applying ethical principles or guidelines upon introducing XAI systems. These principles usually outline the desired or intended outcomes of these respective technologies and system actions, rather than the means to design those systems responsibly and ethically. We contend that the

introduction of ethical principles and guidelines cannot depend upon ad hoc implementations. Principles cannot be defaulted on to be universal solutions, rather their intended use must be specific and context-appropriate within the systems and organizational missions for which they are designed. Therefore, we urge that organizations thoughtfully customize their principles based on a true understanding of their business domain, mission, and culture that conveys the logic and outcome of their principle selections (Ribeiro et al., 2016; Zhang & Li, 2021).

10.1.1. Overview of Explainable AI Principles

Before explaining formal principles in this chapter, we provide a brief conceptual overview of explainability, its importance, dimensions, and established approaches. Our general message is that to realize meaningful explainability, meaningful levels of interpretability and explained understanding must be realized for specific users and specific tasks. While the importance of explainability can seem trivial, it turns out not to be. Consider why human reasoning is explainable when that of advanced intelligent planners often is not. The shallow reasoning represented by logical vectors and propositional logic systems is much easier to reason about than the much simpler operations and state changes of a traditional autonomous planner. Yet, the latter reasoning performs well on important, difficult practical problems.

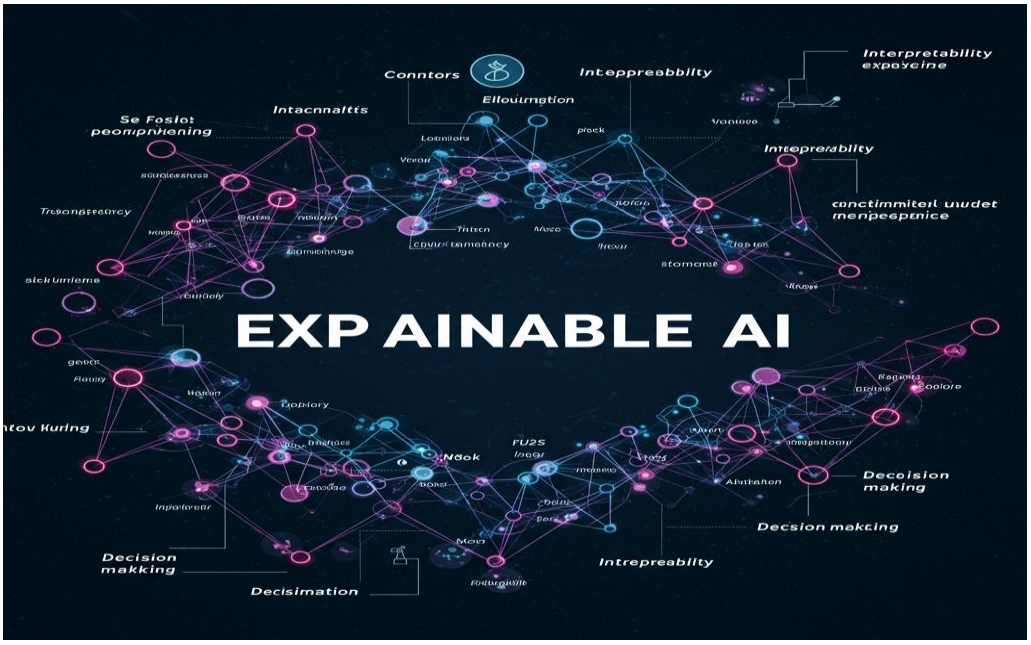

Fig 10 . 1 : Explainable AI

The main sources of difficulty involve the number of concepts involved, their depth or complexity as in multi-layered concepts, and the mental state of the planner. These effects seem to also apply to Markov decision processes and policy explanations based on the policy trajectory and specific policy trajectory distance metrics. We believe that human reasoners have considerable experience and capability in evaluating and explaining the shallow logic level used to describe human reasoning. This factor most likely impacts the effectiveness measure implicit in the principle of explanation self-consistency, which states that the quality of intuitive self-explanatory programs can be great in principle. Evaluation involves predicting and comparing human capability in solving reasoning, perception, narrative, and other problems without the external assistance of machines.

These considerations suggest designing explainable AI for specific users, tasks, and domains in terms of meaningful measures of task difficulty, user characteristics, and AI capability. We devote much of this Tutorial to these issues at varying levels of detail. First, we introduce and elaborate on an overview collection of principles that trace these issues over the entire explanation life cycle. These principles suggest important interpretations of the basic ideas of task, user, and AI agent coding hierarchies and explanation technologies and techniques.

10.2. The Importance of Explainability in AI

Significance of Explainability in Artificial Intelligence

Due to the supervised nature of the algorithm, the need for explanation has received a lot of attention recently. As it has become clear that an increasing amount of industries, such as automobiles, healthcare, and aviation, are relying on algorithms that can not only perform optimally, but, most importantly, can also offer an explanation for their decisions, it has become apparent that one key to trust is explainability. Explainable Artificial Intelligence has responded to this call by finding ways to explain the decisions made by its models. Explanations in general consist of multiple patterns to follow which are provided by models during their training phase. The patterns are learned using a training set, which samples the input feature space as it is linked to the expected output feature. An explanation technique utilizes the patterns learned during the training phase and translates them into useful and interpretable information during the inference phase. While it is expected that, typically, instructions on how to perform a task involve expert background knowledge, a model that learns from experience does not get explicit instructions on how to learn those patterns; as a result, it is difficult to interpret how a model makes its predictions, and that, in turn, is an impediment to achieve a fruitful collaboration with the human expert.

In many cases, expert understanding of the model's decision is not an absolute requirement. In other tasks, such as image classification, animal detection, traffic sign detection, and other similar applications, algorithms are not required to be interpretable, as an error in classification does not involve death or serious injury risks. Despite these, the need for explainability can vary greatly across its use conditions; this means that, depending on the application and the requirements on how models may be exposed to errors, explainability can either be necessary or not.

10.2.1. Significance of Explainability in Artificial Intelligence

Explainable artificial intelligence serves an important function in machine learning by providing more context, insight, and understanding into the often inscrutable workings of complex computational systems. The increasing incorporation of machine learning and related technologies into diverse aspects of day-to-day life raises questions and concerns about accountability, perception of illusory confidence, distribution, bias, model selection, and other issues. The use of deep, complex, and often opaque AI systems hampers understanding and often hides unacceptable biases and weaknesses in operation. The function of explainable AI is to demystify and add interpretability to black-box AI systems in diverse application domains. Explainable AI aims to develop techniques to help with the associated problems and ensure that organizations and individuals can understand the decisions made by AI systems. Addressing the associated challenges is critical in several application domains, including safety-critical systems, such as autonomous vehicles. The opacity associated with black box systems makes it impossible to be sure how such systems will operate in unfamiliar and unanticipated contexts. When comprehensive training for a black box AI system is impossible, the AI system's design cannot incorporate vast experience from prior systems. Explainable AI is also significant in creating human trust and acceptance. In many domains, human users need to trust an AI system before they will be comfortable using it or giving it decision authority. Information retrieval systems, translation systems, decision-making, and assessment systems will interact most effectively with a human user if the user trusts the system when seeking solutions for specific needs.

10.3. Key Concepts in Explainable AI

Despite the increasing importance of explainability in providing AI-enhanced advisory and consulting services, not much guidance is yet available in the scientific literature on the core concepts that characterise explainable AI. Different design aspects have been proposed to enable explanations, including transparency, interpretability, trustworthiness, and others. In this section, we discuss the core concepts of transparency,

interpretability, and trustworthiness, which guided the design of the explainable AI approach that underpins the works presented.

1. Transparency

Transparency can take different forms: through intelligible interfaces and design, through transparency in other interactions with people, or by using intelligible processes that lead to the model's predictions. Transparency can also be based on the motivation for the AI system's predictions, such as by disclosing uncertain or hard-to-interpret instances or exposing model failure cases. It is often associated with inherently interpretable or intrinsically explainable models or systems. These concepts relate to approaches that are considered to enable explanation by design. Some approaches that fall into this category are interpretable models, transparent data processing, or using intelligible transformations in the interpretive process. For instance, symbolic models such as decision trees can qualify as interpretable methods.

2. Interpretability

Interpretability is a characteristic of such processes or models in the context of an unfolding interaction, just in time, where humans attempt to extract the semantics needed to understand and explain the model's predictions or decisions. People use these processes to map the internal representations of AI systems onto concepts in the human cognitive model to understand the system's function. Many explanation approaches that are based on AI-generated interpretations, or what we call interpretation-by-design, fall into this category. Some methods that draw predictions decorate them with attention maps, local explainers, or local faithful approximations. The reliability of any explanation approach is critical in determining whether the terms explanation or understanding are a fair and accurate characterization of what is happening in a given case.

10.3.1. Transparency

Transparency is one of the defining characteristics of XAI technology that enables interactions between the human being and the AI/ML model. Technology Transparency can be defined as the extent to which an intelligent system provides that information. There are two types of technology transparency: design transparency which is defined as the information a system provides about its design, and function transparency which is defined as the information a system provides about its current function. In complex systems, design transparency plays a fundamental role, potentially comprising considerable quantities of engineering and architectural information. In most systems, whose complexity is reflected in intricate interdependencies between system components, it may prove infeasible to lay out all design criteria, guidelines, procedures,

regulations, and design assumptions as well as all design decisions or their rationale. Therefore, it will probably be the task of a system designer to identify and implement a reasonable amount of design transparency. The amount of design transparency is determined by many factors, with the most important being the effectiveness of function transparency. Lowering the level of design transparency demands an enhanced level of function transparency. Function transparency consists of three types: processing transparency, resource transparency, and diagnosis transparency. Processing transparency provides information on the current system operation mode, the status or state of the internal system variables, functions, or resources, and the output of individual processing functions that are part of executing the overall function. Resource transparency identifies high-level components and, for distributed systems, additional nodes as well as represents their specialization and capabilities. Diagnosis transparency includes fault detection and isolation.

10.3.2. Interpretability

Interpretability is among the best-known and widely used concepts in Explainable AI (XAI). It is commonly defined as the degree to which an external observer can understand why a model has made a specific prediction. In other words, given a decision made by the AI agent, how easy is it for an outsider to comprehend the rationales that lead to that specific decision? Interpretability is one of the major concepts justifying the need for AI explainers. It might be hard to believe that a human being can understand decision-making processes at a deeper level than just seeing the results. Nevertheless, the knowledge advanced decision-making systems produce is not an absolute measure of someone's ability. An alarming fact is that the deep learning-based AI agents that have made the biggest revolution in a variety of advanced pattern recognition processes – image recognition, speech decision, and translation between natural languages – have produced intolerable levels of mistakes, often periodically. Even with the surprising success of these systems in terms of the accuracy of the trained workers, such alarming levels of mistakes would normally lead to a search for a solution in classical decision-making systems. The third-party step is considered a decisive factor in some of the most critical applications of AI-based technology, such as medical diagnosis, image, and facial recognition that might lead to arresting, credit scoring, unforeseeable personal insurance premium withdrawal based on own-life prediction, etc.

Decision-making agents apply models – statistical or computational structures that summarize the relations between the explanatory variables and the target variable. The specific value of the target variable for a new entity is predicted using a fixed model and the values of the explanatory variables of the new sample. All agents, no matter how advanced, use models to connect organizational knowledge to their decision-making

capabilities. When models are employed the connection between organizational knowledge and the agents' decision-making process, though hidden and not directly considered, is the sole existing link between the two primarily distanced factors – organizational knowledge and AI agents' output.

10.3.3. Trustworthiness

The notion of trustworthiness can be understood as a saturation of all desirable properties or qualities of a solution, by general orientation of those qualities toward high and suitable values, while being useful as approximated by a worthy vicar, in itself or linked with the context. This idea stems from the following considerations. Transparency and interpretability cannot individually suffice to the system, even if they have a sensible high value. For example, models that are too simple are overly transparent and interpretable but are completely useless. Likewise, a model may be highly complex to understand for a human but utmost precise, accurate, and efficient. This would speak mainly in favor of the accuracy dimension of trustworthiness and the predictive power property. But from a human perspective, the not comprehensible model is not desirable. Now consider the opposite pole of different situations, having either generalized low-complexity and happiness and consequence unhelpful model. In different values of the different factors of trustworthiness, we will have different ranges of usefulness linked with trustworthiness.

In reality, however, this increasing orientation in possible conceivable dimensions may not be competitive with each other. For a high trustworthiness linked with an increased target effect, this may be impossible too. In this sense, trustworthiness is a generalization and a more usable notion. Of course, it is implicitly based also on the other existing concepts like interpretability, coherence, transparency, and global or human-centric over excessive usability… all those notions could be used as dimensions to articulate, implement, and measure the concept of trustworthiness and take all those aspects into account, on the model solution side, as well as the need-side conditions, variables, issues, and opportunities.

10.4. Applications of Explainable AI in Consulting

With the multiple challenges faced by organizations, we believe consultants can greatly benefit from specialized AI tools and solutions. Derived from Adversarial Learning, Explainable AI provides capabilities to both help understand the "how & why" of outcomes identified by general ML algorithms, and create models that are less susceptible to potential adversaries, and discriminative against sub-groups of data. Decision-making processes optimized using Explainable AI algorithms should be more

effective, and also focus attention on key questions that may otherwise be neglected, or even suggested answers sought to be falsified. Explainable AI finds applications in multiple Consulting domains. In Risk Assessment, consulting companies traditionally create internal models to assess the probability of occurrence of events that can endanger company business, including lack of compliance with sanctions regulations, possible damage to corporate reputations, and catastrophic losses from economic crime, among others. Given the critique by Financial Services of these models as Black Box Models, its capacity to assess risks in a more explainable manner should help provide more confidence in the results, increase the speed of model-making in organizations, and provide better input for the remediation actions requested by the regulators when steps have to be taken regarding high-risk customers. In Decision Support Systems, a range of optimization solutions applied in Marketing, Sales & Service functional areas look to balance conflicting objectives with different priorities. Multi-criteria decision-making approaches enable to reach of recommended decisions, using customized models from the user with common innovations in those business areas decided by consensus. In these different functional areas, customer insights require doubtless explanation as being key points of the entire process. It cannot be dependent on algorithms that explain the decisions to be made without their participation. Otherwise, the risks to the accuracy of the results and the effectiveness of the implementations are huge.

10.4.1. Risk Assessment

Over the decades, researchers and business consultants from different disciplines have studied, defined, and proposed various taxonomies of risks of different types. These risk types include financial risks, operational risks, strategic risks, regulatory and compliance risks, information security risks, and reputational risks. In the business context, assessing the organizations' exposure to potential risks is extremely important since poor risk management can lead to the organization's failure or even bankruptcy.

To make risk assessment less burdensome and complicated with some degree of automation, many organizations map indicators from a wide variety of internal and external data sources to individual risk types and perform risk assessment on a scheduled or ad-hoc basis. For example, banks have very mature and well-defined risk management frameworks because regulators monitor their activities very closely using agency-specific guidelines and industry best practices. The main challenge of automating risk assessment is to develop a mathematical framework for time series prediction of risk indicators.

Not only is risk prediction complicated by the current state of the art in time series prediction, but the prediction will fail if the prediction models are not scrupulously validated on fresh test data regularly. The underlying assumption of time series

prediction of risk indicators is that the future will be like the past, especially for well-established economic cycles that have been in existence for decades. With the increasing frequency of external shocks of extreme proportions to established economic cycles such as bank runs, catastrophic natural disasters, economic depressions, and man-made military, political, and economic catastrophes such as wars, civil disturbances, and terrorism, this assumption has become increasingly suspect.

10.4.2. Decision Support Systems

Artificial intelligence (AI) will inevitably play a major role in today's digital transformation. More and more businesses are leveraging intelligent systems for optimizing their internal processes, such as customer service, supply chain management, or fraud detection; as well as their external actions, providing personalized services that meet customer demands in a timely and effective manner. However, social pressures are increasingly demanding a deeper involvement from companies. A lawful and ethically oriented behavior will enhance the value of the organizations, reducing their risks of sanctions as well as reputation damage.

Fig 10 . 2 : AI-Based Decision Support Systems

Despite the fact that company management is mainly concerned with maximizing profits, it must not be forgotten that managers are the agents of the shareholders and thus any decision favoring the interests of the company and society at large will benefit the agents in the long run. There is a conflict of interest between companies and society as a whole. The social dilemma of maximizing individual advantages does not easily find its solution. This conflict allows the right institutions to ensure that the sense of solidarity of the agents is encouraged.

Making the problem of collective choice explicit is an important task. This is one of the various approaches to explaining the behavior of intelligent agents. In particular, decision-making explanation is limited to the area of intelligent systems designed to provide decision support. Many experts in various fields of knowledge use specific intelligent systems to guide them in the development of their activity. Bankers handle credit assessment packages based on both financial ratios and financial data of their customers. Credit risk rating services make a global risk estimation taking into account variables such as economic conditions, political stability, and the credit history of a certain country.

10.4.3. Customer Insights

Explaining the value of present and future customer support is a key concern of large and not such large firms. Abandoned cart clickstream tracking, predictive maintenance, life insurance renewal probability, product launch outcomes, marketing campaign run rates, influencer endorsement efficacy, and more decisions rely on customer insights being easily collected and accounted for. Customers are busy and cannot consume all that a firm willing to promote wants to tell them. Customers are wise and are instinctually reticent about revealing their tag and track invisibly disseminated across the Web. Customers are innocent and sometimes trustworthy when asked personally. And firms are cautious. They want assurance of action and value before overstepping the law and divulging private information too much. As is the case for decision support systems, AI-powered tools can help. Behavioral insight-based decision support systems can help.

The behavioral virtue of data science is that real, objective data can help substantially in the long trek of accumulating insight through human consideration and dialogue. Marketers and users chat at designated spots, sharing comments and content, and queries regarding the widget of concern. Data science can then construct a trail. This probability-based trail, in concert with the internal hashing efforts of the firm, can become a behavioral insight foundation for insight into reducing measurement, texting on the go, targeting surprise messages or enticing content when it counts, and prediction and personalization… when it counts. Explanations can present those events which align

probabilities well, so that insight plus accountability beware. To be fair, internal hash functions often can supply these explanations well without the aid of insight collection.

10.5. Frameworks for Implementing Explainable AI

There are different approaches to implementing Explainable AI (XAI). Such approaches are generally a function of the trade-offs between the need for accuracy, transparency, and explanatory power. In this study, we categorize XAI techniques into three categories as follows. The first category is model-agnostic methods. The second category is interpretable models. The third category is post-hoc explanation techniques. It should be noted the aforementioned categories are paradigm-agnostic, meaning the techniques in the first two categories apply not only to Machine Learning (ML) but also to other paradigms that fall under the broader umbrella term AI.

Model-agnostic methods provide the building blocks for greater transparency and interpretability. Such methods are best suited to scenarios where end-users of AI systems are not experts in the core computational models and need help better understanding the model inputs and outputs. These methods make use of data and/or meta-data to improve model transparency and interpretability. For example, Group/individual feature impact, Interaction Search, Trajectory Search, Input Conditioning, and Adversarial examples methods help improve model transparency and interpretability by use of meta-data. On the other hand, Feature Selection, Feature Manipulation, Subgroup Discovery, Data Documentation, and Examining Training Examples methods help improve model transparency and interpretability by use of data.

Interpretable models enhance transparency and interpretability by exposing simplified versions of the original model, altering its complexity directly. Post-hoc explanation techniques go one step further in terms of user-centricity. Such techniques seek to extract a concisely understandable description of the behavior of complex AI models. In other words, posthoc methods do not focus on the internal workings of a model at a micro-scale but offer a medically plausible explanation based on the overall behavior of the model at a macro scale. Thus, post-hoc explanation techniques rely on the concept of "Interpretable Surrogate" models for achieving their goals.

10.5.1. Model-Agnostic Methods

The increasing demand for explanation methods in machine learning (ML) systems is primarily driven by the need for a better understanding of the reasoning behind the output of such systems. Traditionally, methods in Machine Learning and Artificial Intelligence (AI) were black-box systems that lacked any form of explanation or reasoning behind

their predictions. Due to the increasing applications of machine learning, the practical disregard for explaining results generated by using various ML systems has become a matter of contention. Digital systems are being used now to make decisions that can have life-or-death consequences, whether it be for credit checks, recruitment systems, or other areas like Self-Driving Cars and Image Recognition. In such applications and scenarios, accountability is key to the practical application of AI/ML systems. In this case, explanations help the system owners and the developers to better understand the decision-making mechanisms in the system to identify potential issues in its working and avoid possible faults and errors in the future. Explanations contribute to machine learning accountability – the ultimate goal of explainable AI. This highlights the importance of XAI Systems. The ML community is continuously working towards various levels of explanation that can be provided across various levels of domain complexity.

Such lack of explanation alone may not be the sole reason why explanations are needed in AI systems, in the end, it all comes down to accountability. Our exploration shows that explainers can be categorized based on their capacity on the type of model they provide explanations for. Broadly speaking, model-specific and model-agnostic methods. In model-specific explanation systems, the explainer is capable of providing reasoning based on a pre-trained system that can generate the explanations. In model-agnostic systems, the explainer is a separate system or a model that learns about the relationship between the input and the output of a model and then based on this knowledge, generates explanations for the specific model. In general, model-agnostic systems can provide explanations for a wider array of machine learning systems as compared to model-specific systems with some limitations.

10.5.2. Interpretable Models

Interpretable models are models that are designed to be easily understandable without transforming any information, whether input or output. Highly linear models such as Generalized Linear Models and linear classifiers are widely used interpretable models, as they often serve as the benchmark for model performance. Classical statistical models such as generalized additive models, tree classifiers, and piecewise constant functions have been used in highly structured domains ranging from healthcare to credit risk modeling. These models can leverage human structure in the design of the modeling and approximation, while also fulfilling the goal of transparency.

What is missing from these models of both types is scale; as the size of the data increases and the input and output dimensions increase, we often have to resort to our loss-increasing assumptions and use greater modeling flexibility with black-box predictive models. However, as more industries and more use cases of AI are being discovered,

there exists a desire for more interpretable predictive approaches that are more scalable than known interpretable models. Indeed, even intuiting changes to existing interpretable models or creating a larger space of interpretable models also allows for applications across finance, medicine, and even criminal justice, to name a few. A clear mathematical relationship to the predictions from the input feature dimensions is descriptive enough to be the basis for downstream recommendations.

10.5.3. Post-Hoc Explanation Techniques

Post-hoc explanation techniques allow us to explain the decisions made by complex models and apply them to any model once it has been trained. We divide these techniques into two categories. Local explanation techniques explain a particular decision made by the model, while global explanation techniques help us understand the effect of different features across our entire dataset, identifying global patterns. Many of these techniques are model agnostic and therefore applicable to any model. However, some take advantage of specific properties of certain models to provide better-quality explanations. These techniques provide a unique value because they are relatively easy to implement, easily scale with huge datasets, and often do not require specialized learning.

Local explanations attempt to explain the predictions of complex models for individual instances. These methods are helpful at the local level because they tell us which features were salient at a particular instance-level decision. However, such explanations can be misleading when viewing them globally or at scale because there may be inconsistencies between the explanations provided at different levels of the distribution. These types of techniques provide per-instance explanations that help illuminate decision-making for individual cases. However, interestingly, good models tend to train on decision boundaries without regard to the shape of the decision function. As a result, for many models at certain dataset scales, the local surface of the decision boundary is piece-wise linear, allowing for a simpler model to act as an explanation fit for the entire domain. In other words, many complex models can approximate piecewise linear functions, and therefore, we can approximate them as a simpler model.

10.6. Challenges in Explainable AI

The field of Explainable AI (XAI) has grown in popularity over the last couple of decades, allowing predictions and insights generated by complex AI models to be more easily understood and acted upon. A major challenge of this growth, however, is the actual need for and applicability of these XAI techniques. This chapter will briefly discuss some of the more prominent challenges within this space.

The first challenge is often the complexity of the AI decision model. The more complex an AI model is, the less likely any applied XAI method will help decipher the importance of the input nodes in the actual prediction of the output node. Because popular XAI techniques utilize fluid simulation techniques, they do not scale well with the dimensionality of the input. This is primarily because they reduce the dimensionality of the space of inputs being analyzed and summed into a single “importance” value. Other XAI techniques, such as those based on neural symbolic learning, are still being actively pursued, and it is unclear whether these techniques can truly scale to be applied to complex models, especially considering that many of these techniques are still in a testing, research phase as well.

Another challenge for practical applications of XAI, especially in business, government, or healthcare domains, is that of regulatory compliance. For instance, regulations increase the relevance of XAI: “The controller shall provide … meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject”. With these types of regulations coming from more and more governments worldwide, businesses that do not have an explanation for their model decisions may be subject to significant fines and other forms of regulatory punishment.

10.6.1. Complexity of Models

The term "explainability" is used in very diverse ways. Furthermore, there are a variety of techniques available to achieve "explainable AI". Unfortunately, several of the baseline approaches are not necessarily useful in terms of applicability to a wide range of source model families. These methodologies tend to be confined to performing explainability on "supervised predictive" use cases, delivered by white-box models, or built using very particular "black-box" source model paradigms. On the other hand, human-level comprehension of complex situations may require intricate models that cannot be consigned to a small number of interpretable features or not be summarized in a small number of example cases. Therefore, human-algorithm interaction must be a cooperative effort in these instances.

The complexity and flexibility of the deep learning model families do not lend themselves easily to interpretable summaries of their decision rules or faithful approximations through interpretable surrogates. Despite the many contributions that have been made in this frontline area, and the wide acceptance of the utility of particular methodologies, explaining the decisions of these flexible complex source models is far from a solved problem. Further research can lead to an informed provision of techniques for the understanding and trust of models. In particular, "neural architecture search", is

a promising emerging topic in this area, and in general, the use of principles from cognitive science and neuroscience could lead the field forward.

10.6.2. Data Privacy Issues

Data privacy is often a concern for organizations implementing AI systems. The most powerful techniques, such as multi-layered neural networks, require vast amounts of data for training and effective generalization across populations. Sensitive personal data that is difficult to collect in great quantities often requires additional care to be used, especially when AI systems are being trained for industry applications. Such data may include health information, financial records, education records, and criminal justice records, among others.

Although there are many privacy-preserving techniques proposed, including anonymization to avoid revealing sensitive personal information along with differential privacy and synthetic data generation to avoid leaking information contained in original training datasets, there is generally a trade-off between model explainability and privacy preservation. For example, models with fewer parameters tend to provide better privacy and generalization. However, users may not use these machine learning models because the produced results may not be trustworthy after detecting model bias or other explanation issues. In such cases, a user may tend to distrust invoking these AI models and go back to the previous step, violating the original aim of technology to facilitate and accelerate human work. However, privacy-preserving methods may have the risk of introducing explanation errors into the model. More specifically, some explanations produced by the AI models can potentially compromise privacy.

For example, there is the risk of disclosing sensitive information, such as someone's medical condition through the investigation of a medical prediction model, even if aggregates are introduced to provide explanations. Therefore, careful consideration must be paid to the intended use of the model and its permissible context of use to avoid compromising privacy when using these explanation techniques with predictive models in sensitive domains. Moreover, it poses a greater challenge for regulations, especially data protection directives.

10.6.3. Regulatory Compliance

Much of the news around explainable AI research is focused on algorithm and model development, but as society shapes regulations around the development and deployment of these AI solutions, it is clear that explanation methodologies must also be able to comply with such third-party accountability requirements. Efforts are ongoing to

develop an AI Bill of Rights. The proposed Artificial Intelligence Act is scheduled to shape the regulations around the development and deployment of AI solutions. The AI Act proposal is a risk- and trust-based framework that describes both the obligatory transparency requirements that the operators of high-risk AI systems must meet, employing both technical and non-technical means, and the obligations for the high-risk AI systems themselves. In particular, it states that, when the AI system is used to provide a user with information to decide, the user must be warned if the information is generated with an AI system.

However, much work remains to be done in this space. Currently, neither the United States, nor the European Union or other governments worldwide have set artificial intelligence-related guidelines or laws on what compliance metrics must be checked to make sure models produced by third parties are not harmful, unfair, or privacy-invasive. Instead, great responsibility remains with the companies and institutions deploying such AI models to ensure that they comply (or are being held accountable for what they decide) by being as interpretable, clear, and factual as possible. These AI compliance requirements can be to ensure that the model uses as much information as possible that is contained in the input data and then uses the data to explain its reasons for the particular decision or prediction returned.

10.7. Future Trends in Explainable AI

The emergence of Big Data, which in combination with Cloud Computing has created new opportunities for all fields increasingly dependent on data. AI has been increasingly assuming all daily activities. However, as in any other information-based area, the use of AI for its own sake is not sufficient to necessarily solve problems in the real world. Moreover, organizations are expected to monitor such systems, maintaining things under control. In this sense, Explainable AI is a crucial enabler, and as such, it should be integrated with Big Data. Such integration will be the point in two different directions, along which many future challenges will lie ahead. On the one side, Explainable AI solutions should keep processing, within acceptable times, the vast amounts of data already stored in Big Data repositories. On the other side, extending traditional Big Data with Explainable AI will enable truly driving new insights from the data being processed, capable, for instance, of justifying unexpected results or suggesting new business rules for the analysis of future data.

Natural Language Processing also deserves special attention in future developments of Explainable AI solutions. Natural Language Processing has been rapidly advancing thanks to the emerging transformer models. Nowadays, tasks such as sentiment analysis or recommendation systems, based on understanding and enriching textual information, such as customer reviews, are starting to provide much better results. Related to this,

both billing customer complaints and automatically drafting the answers to those complaints, as well as other daily activities relying on taxonomy and concept creation, are also improving their predictive ability thanks to these new models. The use of transformer models for tasks requiring explanation and understanding of the semantics between different entities of knowledge is contributing to such an increase. The trend is only to increase, although ethical questions and user acceptance will have to be answered.

10.7.1. Integration with Big Data

AI's promise is tied to its unique ability to quickly paint a picture based on large amounts of data designed to answer specific questions. Explainable AI (XAI), with its added capabilities of delivering narrative insights that are trusted and understood, promises to push this ability further and allow a wider set of users to create value using AI. Big data has been around for a long time and its stages and states of development in the field of AI are about the same as those of machine learning or computer vision. Big data-focused clouds, platforms, and products focused on analytics have been churning for a while. AI has only added potency to the mix. The deployment of argumentative analytics, where explanations of analytics lay on, e.g., "What else is like this or unlike this" gives users more intuitive access to results than quantitative insights alone.

The same capabilities can reside inside business management systems or other platforms such as enterprise resource planning systems. An assessment engine has been rolled out that sits inside of an enterprise resource planning system and creates trust scores for all relevant transactions from finance and logistics. These trust scores are based on the risk-versus-opportunity tradeoffs faced by the organization involved because this tradeoff is formalized in the mathematics of the backs of deferred models used for training natural language generation neural networks. These transaction trust scores are then explained to users via expandable argument trees with XAI explaining problem and opportunity statements, possible explanations of the situation, and your role in it. This enables system users to better understand what they are carrying out, make informed decisions about how to handle those transactions, and adjust their behavior accordingly.

10.7.2. Advancements in Natural Language Processing

Natural Language Processing (NLP) has seen immense growth in the past few years largely thanks to the implementation of deep learning techniques to model the representation of words and sequences of words. Language models based on transformer architectures trained on large amounts of text data have pushed the state of the art in a variety of NLP tasks and applications. The availability of trained and open-sourced

trained language models has transformed NLP, making it accessible to a much larger audience of developers with no AI expertise, and paving the way for multimodal models and applications that can encode and relate images and text descriptions, which transcend the borders between modalities. The next advancements in NLP will be research milestones being pursued by the AI community, from more explainable and less biased language models to multilingual, low-resource, and few-shot models. Models with advanced reasoning capabilities for commonsense, multi-hop, and logical reasoning with structured knowledge. Models capable of effectively processing emoji in addition to a variety of modalities beyond text and image, and using other user input formats such as voice with speech-to-text capabilities. Models that can interact with users through dialogues to assist them in accomplishing tasks, or even chatbots that become companions replacing the human interaction of people feeling lonely.

The impact of these advancements will stretch beyond human-centric and multimodal NLP interfaces and user-oriented applications, enabling the industry to automate a wide range of tasks in various domains, such as customer support, code generation, programming tasks, content creation pipelines, compliance, education, healthcare, recruiting, and legal services. Ultimately, NLP will be one of the cornerstones of more intelligent and efficient workplaces. By breaking down the barriers of data, language, resources, and time, intelligent conversational user interfaces will relieve professionals of tedious and time-consuming tasks, enabling them to focus more on the creative and strategic aspects of their work, increasing their satisfaction and efficiency. The combination of NLP with speech recognition, sentiment detection, affective computing, face and emotion image analysis, and conversational AI will intensify this trend in the industry toward automating user and customer-facing activities reliant on language and speech. By making decision-making more efficient and faster through Natural Language Understanding, and decision execution through Natural Language Generation, Human-Machine Collaboration architectures that consider language as the common thread will take off.

10.7.3. Ethical Considerations

At the inception of the AI revolution, there was an implicit understanding that AI would augment and enhance the work we do, not replace it. This high-level endorsement of the AI initiative was without question. The potential of AI was just too enticing. No one at the time thought about how this would be accomplished. Not all use cases are meant to be consumed by machines. As more and more people engaged AI in their business processes, there began dissent and a soaring fear of obsolescence – computers were replacing jobs, and the ramp-up of AI software engineers was limited. With this grass-roots concern came the call for increased regulation.

As AI becomes ubiquitous in our behavior – how we shape and share our realities and how the decisions we make every day affect large-scale data reserves, there is a need to regulate AI so that it acts ethically and to ensure fairness in its decisions. AI learns from history and what has gone before. If there are biases that lead to bad outcomes based on race, gender, or other demographics, there is a very real possibility that those biases will be fed back into society and perpetuated. Also, for AI to act ethically, there needs to be a robust moral principle-based set of values to assess what is ethically right and what is ethically wrong.

To this end, there have been many recommendations recently offered by both government and think-tank initiatives that offer goal-based incentives for research in the area of ethical AI for fairness, with high-impact initiatives that include stakeholder representation and deliberative engagement methods.

10.8. Conclusion

Summary and Final Thoughts on Explainable AI

In this chapter, we discussed the issue of Explainable AI (XAI), including its background, the most popular modal and submodels of the XAI space, and some general models that are often used. After this, we described 14 state-of-the-art explanation algorithms. Based on that, we then provided our resources, syntax, and a tutorial on using our Determining Explainable AI Models Framework. After this, we explained the motivation and scope of this book by introducing some of the many interesting questions related to XAI. We later introduced our taxonomy of XAI and the existing explanation algorithms. We then provided a brief discussion about the validity of explanation algorithms.

Fig 10 . 3 : State-of-the-Art XAI Algorithm Distribution

We concluded our XAI tutorial with a short overview of the ethical, social, and legal aspects of XAI. Our thinking is that researchers developing XAI algorithms need to take into account these aspects laying our developments and future work. The a lack of knowledge about which are the best methodology to develop XAI algorithms for a particular application domain, given the specificities of the decision-maker and that specific domain, the data, and the used model. In addition to that, it is well known that XAI has been used as a generic name for several methodologies that do not provide explanations, only human-interpretable classification. Further, concerning the eXplainable eXact solvers, the ethical concerns, especially around privacy, still apply. Therefore, some additional discussions are needed so that we can clarify the connection between XAI and social good, public interest, privacy, and also the transparency of using human-interpretable models. We hope that by bringing together XAI algorithm development and its implications, we can contribute to the understanding of the subject and help answer some pressing questions the area is confronted with.

10.8.1. Summary and Final Thoughts on Explainable AI

As more and more organizations deploy Machine Learning systems in critical decision-making tasks such as personalized medicine, distribution of financial resources, and hiring decisions, there is a growing demand in the public for transparency in how these decisions are made. This trend is heading to a point where it is almost impossible to build and deploy a Machine Learning system, without some sort of explanation accompanying the automated decision. This process of pulling back the curtain on these Machine Learning systems, so that a stakeholder can gain insight into the reason behind a certain prediction, is known as Explainable Artificial Intelligence. The explanations are useful for increasing trust in the Machine Learning models, as well as giving insights on how to improve the models. Different stakeholders will have different requirements and desires for the explanations, and these demands change depending on the Machine Learning model, and the task it is trying to solve. Moreover, this task is quite difficult, since different explanation methods can also provide varying results. It is imperative to understand what effect the explanation method has on the explanation since different people will have different preferences on how to provide and receive explanations. The current main approaches to explanation generation are various methods that take on a variety of solutions to the presented problem and produce different explanation formats. Finally, we examine the effects of the explanations on the stakeholders and present a survey of fundamental knowledge about Explainable AI that models and modelers should be aware of.

References

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 4765–4774. https://doi.org/10.5555/3295222.3295230

Chen, J., & Zhao, X. (2020). Explainable Artificial Intelligence for Risk Assessment in Financial Consulting. Expert Systems with Applications, 140, 112924. https://doi.org/10.1016/j.eswa.2019.112924

Caruana, R., Gehrke, J., Koch, C., Koch, D., & Lesh, N. (2000). User-Controllable Learning of Classifier Behavior. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 242–251. https://doi.org/10.1145/2982327.2982384

Zhang, Y., & Li, H. (2021). A Survey on Explainable AI for Advisory Systems. Journal of Computational Science, 52, 101396. https://doi.org/10.1016/j.jocs.2020.101396