International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 1, Issue 4, 41-51, 2020
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V1I4P105

Original Article

The Age of Explainable AI: Improving Trust and Transparency in AI Models

Sarbaree Mishra
Program Manager at Molina Healthcare Inc., USA.

Abstract - *Artificial Intelligence (AI) is changing the way healthcare, finance & law enforcement work by making them more efficient & creative and making it easier to make these decisions based on their information. As AI models become more complicated, their decision-making processes become less clear, which hurts trust, accountability & ethical use. Explainable AI (XAI) has come up to help with these problems by making AI systems easier to understand and more open, which helps people understand why certain actions were taken. XAI makes AI's decision-making processes easier to understand by employing these approaches including feature significance analysis, model-agnostic methodologies, interpretable models & visualization tools. These methods make sure that important tasks like medical diagnosis, approving loans & finding bias in law enforcement algorithms are all accurate, fair & easy to understand. XAI gives consumers clear, useful information that helps them make informed decisions with confidence, even if they aren't tech-savvy. This research looks at the main ideas of XAI, including its most important methods, uses & the problems it has to solve to live up to its potential. XAI builds trust in AI systems by making AI models easier to understand. This leads to more widespread use & more responsibility in important areas. As AI becomes better, it will be important to explain things in order to deal with moral difficulties, reduce bias & make sure that people follow the rules. This will help people utilize AI technology in a more responsible & sustainable way.*

Keywords - *Explainable AI, Trust, Transparency, AI Models, Interpretability, Accountability, Machine Learning, Decision-Making, Ethical Standards, Regulatory Requirements, High-Stakes Applications, Healthcare, Finance, Legal Systems, Confidence, Accessibility, Reliability, Explainability, Fairness, Bias Mitigation, Model Interpretability, Model Transparency, Predictive Models, Black-Box Models, White-Box Models, Feature Importance, Algorithmic Accountability, Responsible AI, Human-AI Collaboration, Risk Assessment, Model Validation, Auditability, Explainability Tools, Trustworthy AI.*

1. Introduction

Artificial Intelligence (AI) is becoming an important part of daily life. It has changed fields including healthcare, banking, transportation & education. Its ability to quickly analyze huge amounts of information, find patterns & make decisions has helped predictive analytics, natural language processing & autonomous systems make progress. However, along with these changes, a big problem has come up: AI models are commonly called "black boxes" because they are hard to understand. Some AI systems are hard to understand because they can make highly accurate predictions or decisions, but the reasons & logic behind their outputs are still unclear, even to the people who make them. Lack of openness creates moral, legal, and practical problems, especially in situations where trust, fairness, and responsibility are very important.

For example, when an AI model turns down a loan application, gives a medical diagnosis, or makes decisions in the criminal justice system, people who use it and those who make rules want to know why. If these AI systems aren't clear about how they work, people could think that their results are random, biased, or unfair, which would make them less likely to trust them. Explainable AI (XAI) wants to solve these kinds of problems by coming up with ways to make AI systems easier to understand & more open. XAI distinguishes itself from these traditional AI by aiming to harmonize expected accuracy with human understanding. Traditional AI emphasizes speed & accuracy. XAI improves understanding of AI algorithms, allowing enterprises to comply with these rules, detect and correct biases & increase decision-making.

1.1. The Rise of Artificial Intelligence

Artificial intelligence has evolved from an abstract concept to a concrete reality and is now used in almost all these industries. Substantial progress has been made in areas like speech recognition, image processing & these predictive analytics due to its capacity to replicate human cognitive functions & also decision-making. As AI becomes more ubiquitous, concerns over its

reliability, fairness & the accountability have escalated. These difficulties intensify when AI decisions substantially impact humans' lives.

1.2. The Black Box Conundrum

The "black box" phenomenon emerges from the intricacy of some AI models, particularly those using DL. Neural networks have considerable capability, however they rely on their millions of components that operate in the conjunction, sometimes in obscure ways. These models may be quite more accurate, but their lack of transparency makes it hard to trust their results, especially in important situations. This lack of transparency has become a major obstacle to broader use & raises moral and also legal questions.

1.3. The Rise of Explainable AI (XAI)

The black box dilemma revealed how crucial it is to have models that are both accurate & simple to comprehend. Explainable AI was designed to fix this. XAI's mission is to make AI systems easy to comprehend by creating these tools & methodologies that help people understand how AI systems make choices. Explainability is particularly very crucial in areas like healthcare, finance & criminal justice, where fairness and accountability are very vital.

Fig 1: Explainable AI

2. Foundations of Explainable AI

Explainable AI (XAI) is a field that works to make sure that AI systems work in a manner that people can understand. The main goal of XAI is to build trust, accountability, and openness in AI systems such that they are not just accurate but also easy to understand. This section explains the ideas, parts & steps that make up XAI, giving a structured look into what it is based on.

2.1. What is AI that can be explained?

Explainable AI means that AI systems can explain their decisions & actions in ways that people can understand. Explainable AI (XAI) wants to make it clear how regular AI models work, particularly these deep learning systems, which are frequently seen as "black boxes."

2.1.1. Why Explainability Is Important

There are several reasons why explainability is important.

- Trust: Users are more likely to trust AI systems when they understand why certain decisions were made.
- Accountability: Companies in important fields like healthcare & finance need to make sure that what they do makes sense and can be looked at.
- Ethical AI: Explainability makes it easier to find & fix biases, which makes sure that AI outputs are fair & also reasonable.
- Regulatory Compliance: Governments and businesses are increasingly demanding that AI be open and clear, which makes explainability a legal requirement.

2.1.2. Main Problems with Getting Explainability

Getting explainability is frequently very hard, even if it's really important. The main problems are: how complicated the models are. Deep neural networks & many other advanced AI models have complex topologies that are very hard to understand. Agreements Finding the right balance between accuracy & the interpretability: making models simpler to make them clearer may make them very less accurate in predicting things. Requirements for a certain industry: Different companies require different kinds of explanations, which makes it harder to be consistent.

2.2. Basic Ideas behind Explainable AI

There are a few basic rules that govern the creation and usage of XAI.

2.2.1. Understandability

Interpretability is how well a person understands how the inputs & outputs of an AI model are related. There are two main ways to achieve this:

- Intrinsic Interpretability: Using models that are easy to understand, such as linear regression or decision trees.
- Post-Hoc Interpretability: Using techniques like feature significance or SHAP (Shapley Additive Explanations) to make sense of these complicated models.

2.2.2. Clarity

For openness, AI systems need to be able to explain how they work, why they do what they do, and what their limits are. For example: Algorithmic Transparency: Making clear how a model makes its decisions.

- Clarity in the process: Making clear how the system changes and handles data.

2.2.3. Real Functionality after That

AI systems must provide explanations that accurately reflect how they make decisions. People may become confused & lose trust if they don't have enough information.

2.3. Main Methods Used in Explainable AI

Different methods make it easier to understand AI & they are generally tailored to the model's complexity and use.

2.3.1. Techniques that don't rely on models

Several AI frameworks can function with model-agnostic approaches:

- LIME (Local Interpretable Model-Agnostic Explanations) provides clear explanations of what model predictions mean for certain inputs.
- SHAP breaks down forecasts into the contributions of each feature, giving all models a standard way to do things.

2.3.2. Unique Methods Used by the Model

Some models work better with some methods of explainability than others.

- Determination Trees: Easy to understand since their visual structure shows clear paths for making these decisions.
- Linear Models: Make it easier to understand coefficients & show how the input's traits affect the output.

2.4. The Future of Explainable AI

XAI is continually changing & the things that affect its path are always changing:

- Design with people in mind: Usability will be the most important thing for future XAI systems. They will make sure that explanations match the demands & skills of users.
- Hybrid Models: Combining simple models with these complicated parts to find a middle ground between accuracy & ease of understanding.
- Standardization Initiatives: Setting industry standards and benchmarks for XAI so that all applications work the same way.

XAI might bring together machine intelligence & human understanding by addressing basic problems and following these specific rules. This would lead to AI systems that are strong, dependable, and clear.

3. Techniques for Explainability

People have long criticized AI models, especially complex ones like deep learning and ensemble methods, for being "black boxes" that are hard to understand. Explainable AI (XAI) wants to fill this gap by making it clear how these models make decisions. This section outlines important tactics for explainability, focusing on their structure & how they might be used in technology.

3.1. Techniques for explaining things after the fact

Post-hoc methods explain how an AI model acts after it has been trained & put into use. These tactics work well to explain these predictions and build trust among all parties involved.

3.1.1. Importance of Features

Feature significance analysis finds the dataset's features that have the most effect on the model's predictions. Methods include:

- The Importance of Permutation: This means changing the values of a feature & seeing how it affects the model's performance. A drop in accuracy shows how important the attribute is.
- Specific Approaches to Models: Decision trees and random forests automatically provide feature significance ratings based on how often a feature is used to partition data during training.

3.1.2. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a model-agnostic strategy that makes individual predictions clearer by using a simpler, easier-to-understand model as a local approximation of the original model. The steps are as follows:

- Taking samples from data points that are close to the event being studied.
- Training a linear model to act as the sophisticated model does in this little area.
- Seeing the feature weights in the reduced model helps better understand the forecast.

3.2. Ways to Make Things Easier to Understand

One of these tactics is to create models that are easy to understand on their own, which ensures transparency without the need for further explainability steps.

3.2.1. Systems That Follow Rules

Rule-based systems, which are common in these expert systems, employ "if-then" rules to make decisions. They are easy to understand since the reasons for each option are clearly explained. Still, these systems may have trouble with scalability and adaptability.

3.2.2. Logistic Regression and Linear Models

Some of the simplest interpretable models include linear models and logistic regression. Their coefficients make it very evident how the input features are related to the output.

3.2.3. Trees of Decision

Decision trees show how decisions are made in a visual way. Each node represents a feature split, which makes it easier to follow the path of a choice. They may not be able to predict these things as well as more advanced models, but they are great at explaining things.

3.3. Approaches that don't depend on a model

You may apply model-agnostic approaches with any other AI architecture, which makes it easier to understand different frameworks.

3.3.1. PDPs, or Partial Dependence Plots

Partial Dependence Plots (PDPs) show how a feature relates to the model's predictions by averaging the effects of many other elements. This strategy is very useful for understanding nonlinear relationships in more complex models.

3.3.2. SHAP, or SHapley Additive exPlanations

SHAP is a technique from game theory that measures how much each feature adds to a specific prediction. SHAP gives a systematic way to understand these model results by giving "credits" to features based on how much they provide to the model.

- The main benefit is that you may get a global view by combining these individual SHAP values.

- Visualizations, including summary plots & dependent plots, can show how features are related to one another.

3.4. Understanding By using visualization

Visualization techniques are powerful tools for making these complicated model behaviors easier for more people to understand.

3.4.1. Thermal Maps and Grouping

Heatmaps and clustering visualizations let you see patterns & the importance of these features in tabular information. These tools help those who are interested in the data understand the big picture and how it affects the model.

3.4.2. Maps of Importance

Saliency maps show which sections of an input (such pixels in an image or words in a phrase) have the most effect on the model's prediction. This technology is widely utilized in these applications for computer vision & NLP.Grad-CAM (Gradient-Weighted Class Activation Mapping) produces heatmaps for convolutional neural networks (CNNs) that show which parts of an image have an effect on the model's decision-making process.

4. Real-World Applications of XAI

Explainable AI (XAI) is changing these industries quickly by making it possible for people to understand complex ML techniques. Explainable AI (XAI) lets people understand and trust AI outcomes, whereas these traditional AI systems are typically "black boxes." This article looks at the actual world uses of XAI, which are grouped into important domains & use cases.

4.1. Health Care

AI has been applied in the healthcare field for diagnosis, treatment recommendations & these improving operations. XAI is important for building trust in their AI systems among doctors and patients.

4.1.1. Tools for Diagnosis

AI models have shown great promise in finding these diseases including cancer, diabetic retinopathy & many heart problems. XAI makes these systems better by explaining why they get their diagnostic results. A software that uses XAI could be able to find the specific patterns in medical images, such abnormal cell structure, that led to a cancer diagnosis. This transparency makes physicians more likely to accept AI guidance, especially when they have to make these important decisions.

4.1.2. Personalized Treatment Plans

Explainable Artificial Intelligence (XAI) is used to back up therapeutic suggestions by describing why they were made. For instance, an AI system that suggests chemotherapy may back up its suggestion by looking into the patient's medical history, genetic markers & how they reacted to similar treatments in the past. By making these findings public, XAI lets doctors check and improve the AI's suggestions.

4.2. Money Issues

XAI is an important part of these AI applications in banking since trust & compliance are so important.

4.2.1. Finding Fraud

AI models are quite good at spotting more unusual patterns in these transactions, which is a key symptom of fraud. Explainable Artificial Intelligence (XAI) helps financial companies understand why certain transactions are marked as suspicious. An XAI system may explain that a flagged transaction was different from how the user usually spends money or came from a place with a lot of crime. For compliance teams to do further research, this transparency is really very important.

4.2.2. Checking Credit

AI is widely employed to check creditworthiness, however algorithms that are not clear might lead to biases or wrong credit denials. XAI helps lenders explain their credit decisions, such as why an application was approved or denied. For example, it can mention low income or an inconsistent history of paying back loans as problems that need to be addressed, while also pointing out ways to improve the score.

4.2.3. Trading Using Algorithms

AI computers look through huge datasets to locate good trading chances. XAI gives traders information on why a model chose a given investment strategy. For example, it may explain that a proposal came from looking at previous trends, the market's ups

and downs, and most current economic statistics. This lets traders figure out how reliable the AI's predictions are & keep their faith in these automated solutions.

4.3. Legal and Compliance

The legal field & regulatory bodies need transparency to make sure that AI systems work within the law & in a way that is moral.

4.3.1. Looking at Contracts

AI-powered tools for analyzing contracts can find more dangerous clauses or highlight these differences. XAI makes these tools better by explaining why a language is regarded harmful, utilizing legal precedents or common practices in the business. This makes it easier for lawyers to quickly & accurately review contracts.

4.3.2. Following the Rules

Many industries have to follow these strict rules set by the government. XAI makes it easier to follow the rules by explaining how AI models understand them. XAI systems may help banks make decisions about anti-money laundering (AML) processes by finding more patterns of suspicious activity & linking them to rules that must be followed.Retailers are utilizing AI more and more to make shopping better for customers. XAI makes these systems more open, which makes customers more confident.

4.4. Endorsements of Products

AI-powered recommendation systems provide users goods based on what they like & what they've looked at before. XAI lets these algorithms explain their ideas. For example, they may say that a suggested product is popular with many others who have similar interests or that it improves an item that the user already has. This builds trust and encourages others to become involved.

4.4.1. Looking at what customers say

Retailers typically employ AI to look at customer feedback & find many ways to make these things better. Explainable Artificial Intelligence (XAI) makes the results of sentiment analysis clearer by showing how particular phrases or patterns in the input changed the AI's judgment. This lets businesses employ AI discoveries to make more targeted plans.

4.5. Systems that work on their own

XAI is very important for keeping people safe & accountable in these fields like transportation and robotics.

4.5.1. Cars That Drive Themselves

Self-driving cars rely on more complex AI systems to steer and make split-second decisions. Explainable Artificial Intelligence (XAI) can inform you why a car did something, such as suddenly slowing down or taking a different path. For instance, it can include sudden obstacles, changes in traffic patterns, or weather conditions as possible reasons. These explanations are important for building public trust and understanding of autonomous technology.

4.5.2. Automation in the Factory

Industrial robots typically make production floors more efficient. Explainable Artificial Intelligence (XAI) helps operators understand how robots make decisions, such as why a robot could reorganize manufacturing lines or find a likely problem in a product. This makes it easier for people and robots run by AI to work together without any problems.

5. Challenges in Achieving Explainability

Explainable AI (XAI) is becoming a major issue for researchers, programmers & also policymakers. There are many possible benefits to making these AI systems more open, but getting them to be explainable is not easy. These problems have to do with technology, operations & ethics, which shows how hard it is to make complex AI systems understandable without losing their performance.

5.1. How complicated modern AI models are from a technical point of view

Deep learning models and other modern AI systems work like "black boxes," making it impossible to understand how they make decisions.

5.1.1. Give and Take between Accuracy and Understandability

It might be hard to get great accuracy while still giving these explanations. When it comes to big, messy data like photographs or text, complex models frequently work better than simple, easy-to-understand ones. Developers that care about performance may have an issue with this since making these models easier to understand may make them less accurate at predicting.

5.1.2. Mechanism that is hidden The Core of Deep Learning

It may be hard to describe how deep learning models, such as convolutional neural networks (CNNs) & more recurrent neural networks (RNNs), function because they use layers of abstract computations. These layers look at data with a lot of dimensions & their outputs are generally shown as probabilities or changes to features. It is hard to explain why these systems make the decisions they do since there are no apparent cause-and-effect relationships between them. For example, figuring out why a neural network marked a transaction as fraudulent would require breaking down millions of computations, which is not something that people can easily understand.

5.2. No Standardized Metrics for Explainability

Another big problem is that there isn't a widely accepted way to measure explainability.

5.2.1. Subjectivity in how things may be understood

Different groups have quite different ideas on what an "explanation" is. An end user could just require a simple summary, whereas a data scientist would want detailed information on model properties and weights. This subjectivity makes it hard to set up uniform ways to quantify the explainability

5.2.2. Different Needs of Stakeholders

Many people utilize these AI systems, such as developers, businesses, regulators & end users. Each group has its own needs & the requirements for explanations. It is quite hard to combine these different needs into one system since it may need to have several levels or types of interpretability, which might make the design process more harder.

5.2.3. Problems with measuring explainability

Unlike criteria like accuracy, precision, and recall, explainability doesn't have a clear-cut way to be measured. How do you decide whether an explanation is good enough? When people try to quantify explainability, they generally use qualitative ratings, which may be unreliable and change depending on the situation.

5.3. Problems with ethics and rules

Explainability is closely related to moral concerns and following the rules set by the government.

5.3.1. Following the Rules

Regulations like the EU's General Data Protection Regulation (GDPR) stress the "right to explanation," which forces businesses to make their AI systems easier to understand. It is hard to turn the requirements of the law into these technical solutions that can be used. To be compliant, you frequently have to find a balance between being able to explain things & other things like protecting their information & making sure the system works well.

5.3.2. Making sure fairness and reducing bias

AI systems may provide unfair or discriminatory results if they are biased, which is usually made worse by not being able to explain why. For instance, if an AI model turns down a loan application, the user may ask for an explanation to make sure the decision was fair. When people aren't upfront about their decisions, it makes it harder to find & fix biased decision-making processes, which may lead to more ethical problems and even legal problems.

5.4. Things that make it hard to put into practice

Even while there are ways to make things easier to understand, putting them into practice in real-world systems is still hard.

5.4.1. Problems with Scalability

Most of the explainability tools available today, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), need a lot of resources. These methods are generally not useful for these big systems or applications that need actual time predictions since they might add a lot of time to processing.

5.4.2. Finding a Balance between Privacy and Openness

There is a trade-off between being open & protecting intellectual property or private information in these certain fields, such as banking & healthcare. Offering detailed explanations might accidentally expose private algorithms or sensitive information, which would create a contradiction between openness & secrecy. To stay in more compliance and keep their competitive edge, businesses must carefully manage this balance.

5.4.3. Working with old systems

Many companies still utilize these previous systems that don't come with built-in explainability. It could be hard and costly to add modern explainability tools to these systems. Also, these kinds of interfaces may need a lot of work to change the way things are done now and the way data flows, which might mean that they take longer to set up and cost more to run.

6. The moral and social effects of explainable artificial intelligence

The need for Explainable AI (XAI) has grown clearer as we depend more & more on AI to make these kinds of judgments. The main purpose of XAI is to make the AI models more reliable and open, but it also has social & moral repercussions that go beyond the technology. This section talks about these effects and stresses the need of moral responsibility, fairness in society & widespread trust in these AI systems.

6.1. Moral Responsibility in AI

AI systems should be in line with society norms, moral values & well-known ethical standards, which is what ethical responsibility in these AI means.

6.1.1. Stopping Bad Use

It is less probable that AI technology that is open & clear will be utilized for more harmful reasons. Explainable AI lets us check to see whether AI systems are following certain rules of conduct. This stops people from using them to snoop, transmit faulty information, or behave in a manner that is unfair to others. XAI's moral rules make sure that technology is used for good & not for negative purposes.

6.1.2. Making choices is your job.

It's hard to know who is to blame when their AI systems make mistakes or demonstrate bias, which is a huge concern. Explainable AI deals with this problem by making it easier to understand how choices are made. It is simpler to hold these creators, operators, or organizations accountable when the AI systems are open. People are more accountable and more likely to obey the rules when they know what's going on.

6.2. Fairness and Less Bias

AI systems usually pick up on the biases that are present in the information they are trained on. This prevents things like employment, credit distribution & law enforcement from becoming the same. Explainable AI may help discover & rectify these biases, which would make things more fair and just.

6.2.1. How to Find Algorithmic Bias

XAI enables people who have a stake in the outcome to understand how models look at the information & make decisions. With this knowledge, it is easier to uncover the latest biases in the model's logic or training information. For example, if an AI system always says no to loans to certain groups of individuals, XAI methods may be able to establish that the judgments are unfair.

6.2.2. Making sure that choices are fair

Explainability ensures assurance that AI choices are right & can be backed up. XAI systems make it simpler for people to challenge unjust results by making the process of making decisions more apparent. Candidates who are turned down for a job by an AI-powered recruiting system may get specific explanations for their rejection & tips on how to improve, which might make them feel like they weren't treated unjustly.

6.2.3. Moving ahead with data practices that involve everyone

Explainable AI highlights how crucial it is to have datasets that are more varied & also representative. Engineers may make sure that these AI systems don't depend too much on data sets that promote social disparities by working out which factors impact model predictions. To get the same findings for everyone, it's important to use these data techniques that are open to everyone.

6.3. Building Trust in AI Systems

AI systems need to be more reliable in order to be used in these sensitive areas like healthcare, criminal justice & finance. Explainable AI is necessary to build this trust.

6.3.1. Making it easier for stakeholders to work together

Explainable AI makes it easier for developers, regulators & end-users to work together. Authorities may check to see whether people are following ethical & legal rules in systems that are open and honest. Consumers can also provide feedback on how to make things better. This cooperative approach makes sure that AI technology develops in a safe way.

6.3.2. Building Trust in Users

Users trust AI more when they understand how it makes decisions. When consumers and medical professionals can comprehend why a diagnosis or treatment prescription was made, they are more likely to adopt these diagnostic AI systems. Explainability links complex algorithms with end-users, which builds trust & also acceptance.

6.4. Ethical Problems with Global Execution

The deployment of AI throughout the globe raises greater ethical issues, particularly because of differences in culture & the law. We can use explainable AI to solve these types of difficulties. Because of differences in the culture, AI systems built in one place may not obey the moral norms of another place. Explainability allows people to alter AI models to match local moral standards without hurting their integrity. It also makes sure that regional regulations are followed, like GDPR in Europe, which puts transparency & accountability at the top of the list when it comes to automating these decision-making.

6.5. The Future of Ethical AI with Explainability

As technology becomes better, the moral & social effects of XAI are also becoming better. To make sure that AI is used responsibly in the future, organizations need to focus on both explainability & their creativity. Adding ethical concerns to the creation of AI can ease people's anxiety & give them a sense of trust.

7. Conclusion

The time of explainable AI (XAI) is a key turning point in the development & usage of AI. XAI is different from traditional black-box models that are very frequently hard to understand. It aims to make the decision-making processes of AI systems clear & easy to understand. This change is necessary to build trust, particularly in these sensitive areas like healthcare, finance & law, where fairness and accountability are very important. XAI makes AI systems easier to understand, which helps bridge the gap between technical complexity & human understanding. This builds confidence among users and the stakeholders. But getting to explainability is still quite hard. Finding a balance between model performance and interpretability takes a lot of work, since simpler models are generally easier to understand but may not be able to handle their certain difficult tasks.

Also, XAI has to deal with the possibility of systemic biases to make sure that the explanations are both accurate & fair. Even with these problems, XAI has the capacity to change everything. Making AI systems more open and clear helps businesses get humans & machines to work together better, which guarantees fair, moral & effective decision-making. Explainable AI makes it easier to use AI more widely by easing worries about bias and accountability in the law and society. As researchers and practitioners improve XAI methods, their impact will go beyond just making technology better. It will change how people think about AI by proving that smart systems can work ethically and honestly. In the end, explainable AI goes beyond just understanding how AI works; it means building systems that fit with human values to make sure that AI is a good thing in society.

References

[1] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

[2] Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 6(1), 2053951719860542.

[3] Allam, Hitesh. *Exploring the Algorithms for Automatic Image Retrieval Using Sketches*. Diss. Missouri Western State University, 2017.

[4] Samek, W. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

[5] Immaneni, J. (2020). Using Swarm Intelligence and Graph Databases Together for Advanced Fraud Detection. *Journal of Big Data and Smart Systems*, *1*(1).
[6] Khedkar, S., Subramanian, V., Shinde, G., & Gandhi, P. (2019). Explainable AI in healthcare. In Healthcare (april 8, 2019). 2nd international conference on advances in science & technology (icast).
[7] Jani, Parth. "UM Decision Automation Using PEGA and Machine Learning for Preauthorization Claims." *The Distributed Learning and Broad Applications in Scientific Research* 6 (2020): 1177-1205.
[8] Patel, Piyushkumar, and Disha Patel. "Blockchain's Potential for Real-Time Financial Auditing: Disrupting Traditional Assurance Practices." *Distributed Learning and Broad Applications in Scientific Research* 5 (2019): 1468-84.
[9] Hagras, H. (2018). Toward human-understandable, explainable AI. Computer, 51(9), 28-36.
[10] Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019, May). Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-15).
[11] Manda, Jeevan Kumar. "Cybersecurity strategies for legacy telecom systems: Developing tailored cybersecurity strategies to secure aging telecom infrastructures against modern cyber threats, leveraging your experience with legacy systems and cybersecurity practices." *Leveraging your Experience with Legacy Systems and Cybersecurity Practices (January 01, 2017)* (2017).
[12] Fox, M., Long, D., & Magazzeni, D. (2017). Explainable planning. arXiv preprint arXiv:1709.10256.
[13] Arugula, Balkishan, and Sudhkar Gade. "Cross-Border Banking Technology Integration: Overcoming Regulatory and Technical Challenges". *International Journal of Emerging Research in Engineering and Technology*, vol. 1, no. 1, Mar. 2020, pp. 40-48
[14] Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6-1.
[15] Mohammad, Abdul Jabbar. "Sentiment-Driven Scheduling Optimizer". *International Journal of Emerging Research in Engineering and Technology*, vol. 1, no. 2, June 2020, pp. 50-59
[16] Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019 (pp. 1078-1088). International Foundation for Autonomous Agents and Multiagent Systems.
[17] Manda, Jeevan Kumar. "Cloud Security Best Practices for Telecom Providers: Developing comprehensive cloud security frameworks and best practices for telecom service delivery and operations, drawing on your cloud security expertise." *Available at SSRN 5003526* (2020).
[18] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.
[19] Jani, Parth. "Modernizing Claims Adjudication Systems with NoSQL and Apache Hive in Medicaid Expansion Programs." *JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE)* 7.1 (2019): 105-121.
[20] Immaneni, J. (2020). Building MLOps Pipelines in Fintech: Keeping Up with Continuous Machine Learning. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *1*(2), 22-32.
[21] Nookala, G. (2020). Automation of privileged access control as part of enterprise control procedure. *Journal of Big Data and Smart Systems*, *1*(1).
[22] Fellous, J. M., Sapiro, G., Rossi, A., Mayberg, H., & Ferrante, M. (2019). Explainable artificial intelligence for neuroscience: behavioral neurostimulation. Frontiers in neuroscience, 13, 1346.
[23] Shaik, Babulal. "Network Isolation Techniques in Multi-Tenant EKS Clusters." *Distributed Learning and Broad Applications in Scientific Research* 6 (2020).
[24] Preece, A. (2018). Asking 'Why'in AI: Explainability of intelligent systems–perspectives and challenges. Intelligent Systems in Accounting, Finance and Management, 25(2), 63-72.
[25] Patel, Piyushkumar. "The Evolution of Revenue Recognition Under ASC 606: Lessons Learned and Industry-Specific Challenges." *Distributed Learning and Broad Applications in Scientific Research* 5 (2019): 1485-98.
[26] Sai Prasad Veluru. "Real-Time Fraud Detection in Payment Systems Using Kafka and Machine Learning". *JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE)*, vol. 7, no. 2, Dec. 2019, pp. 199-14
[27] Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. arXiv preprint arXiv:1907.12652.
[28] Manda, Jeevan Kumar. "AI And Machine Learning In Network Automation: Harnessing AI and Machine Learning Technologies to Automate Network Management Tasks and Enhance Operational Efficiency in Telecom, Based On Your Proficiency in AI-Driven Automation Initiatives." *Educational Research (IJMCER)* 1.4 (2019): 48-58.

[29] Ahmad, M. A., Eckert, C., & Teredesai, A. (2019). The challenge of imputation in explainable artificial intelligence models. arXiv preprint arXiv:1907.12669.

[30] Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint arXiv:1902.01876.