Infosys®
Navigate your next

VIEWPOINT

UNLOCKING THE BLACK BOX WITH EXPLAINABLE AI

As artificial intelligence (AI) systems gain widespread adoption in various industries over the next several years, the need for explainability and trust in the decisions made by these systems is increasing. Explainable AI (XAI) addresses the major issues that hinder our ability to fully trust AI decision-making — including bias and transparency — and through its application, can help ensure better outcomes for all involved.

Infosys® | Knowledge Institute

Unlocking the black box with explainable AI

The next decade will be marked by rapid advancements in AI and a corresponding effect on the global economy. By 2030, AI is expected to raise the global GDP by an estimated $15.7 trillion.[1] Research data shows that the number of organizations that adopted AI has grown by 270 percent since 2015. In the past year alone, that number has tripled.[2] However, there is still concern in the business community as to whether humans can be confident about the trustworthiness of AI recommendations.

This concern has created a need for transparency in machine learning, which has led to the growth of XAI. This new field of AI brings accountability to the space to ensure that AI benefits society instead of harming it. Two of the biggest issues that XAI addresses are bias in AI systems and opaque AI decision-making.

Dealing with bias in AI decision-making

XAI will be critical in helping with the bias inherent to AI systems and algorithms, which are programmed by people whose backgrounds and experiences unintentionally lead to the development of AI systems that exhibit bias. Add in the fact that the teams who shepherd AI systems from inception to deployment don't represent society at large, and you've got a recipe for skewed datasets and algorithms with inherent systemic biases.

One such example comes from Amazon, which utilized AI in an experimental hiring tool that rated job applicants on a scale from one to five, with five being the best qualified.[3] It wasn't long before the company spotted bias in its system: The tool was rating men more highly for software developer jobs than women. Why? Because Amazon's system had observed the resumes submitted to Amazon over the past decade — a dataset largely comprised of male applicants — and used that filter to judge candidates.

Because of this flawed dataset, Amazon's system had developed a bias toward men. Resumes that included women were rated lower, and two applicants from women's colleges were downgraded just because of where they went to school. Amazon changed its system to make it gender-neutral, but as this program evolves, there is no way to guarantee that future systemic biases won't develop.

Offering a new level of transparency with AI systems

For years, AI systems were black boxes with little insight available into why the machine reached the conclusion it did. IBM Watson, which beat some of the best human players on Jeopardy, was marketed to hospitals as a new partner for detecting cancer.

When placed within the oncology department at hospitals, IBM Watson did not succeed in its new role. The doctors and patients were unable to trust the machine at each stage of consulting and treatment. Watson wouldn't provide the reasons for its results, and when its results agreed with the doctor's, it couldn't provide a diagnosis.[4]

In other words, IBM Watson lacked transparency. However, the days of black box decision-making by AI systems appear to be coming to an end. New laws, including Europe's General Data Protection Regulation, or GDPR, are bringing the "right to explanation" front and center. XAI provides better mechanisms to comply with accountability requirements within the organization for auditing and other purposes. Thus, there is better adherence to regulatory requirements like GDPR.

Humans must be able to trust AI decisions

For humans and AI systems to work collaboratively, which is expected to happen in the next two years, there needs to be a high degree of trust in the decisions being made by these machines. For that trust to develop, XAI must be applied to answer a few key questions about how AI systems make decisions:

1. For specific predictions or decisions, why was that conclusion reached?
2. Why was an alternative decision not made instead?
3. How do we measure the success or failure of an AI system?
4. As humans, when do we have enough confidence in AI systems to trust them?
5. How can AI systems correct errors that arise?

To better understand specific gaps around our knowledge of AI decisions and how XAI can be applied to remedy those issues, we will examine two industries: financial services and healthcare. In both industries, the financial stakes are high, and as we'll see, XAI can be used to better understand AI decisions in both industries.

The applications of XAI in financial services

By 2020, online fraud detection spend is expected to reach $9.2 billion.[5] Fraud is happening faster and is more

sophisticated than ever, making patterns difficult for financial crime teams to detect, especially among evolving customer behaviors.

XAI can help companies understand how algorithmic models detect potentially fraudulent transactions by adjusting on the basis of changing customer spending habits and life circumstances, thereby leading to a reduction in fraudulent transactions, minimizing false fraud flags and significantly reducing expenses.

> XAI can help companies reduce cyber frauds and related expenditure by providing them an in-depth understanding of algorithmic models

The global AI in the fintech market is set to grow from $1.4 billion in 2018 to $5.6 billion by 2024 at a compound annual growth rate (CAGR) of 26.2 percent between 2019 and 2024.[6] As more financial services and fintech startups leverage AI for various applications, the need for transparency into the decisions made by AI systems will only increase.

For example, if an AI-powered system denies a person's loan application, bank executives should have the ability to review the AI's decision-making step-by-step to determine exactly where the denial occurred, as well as why the loan was denied.

To offer another example, consider an AI system determining premium charges for a car insurance policy. That system should be able to explain what factors — be it accident history, car type, mileage, etc. — contributed to the decided-upon premium. Based on those factors, the system should then be able to offer personalized recommendations to reduce that premium charge (for example, no speeding tickets for a year).

The applications of XAI in healthcare

As AI has spread across various industries, one of the areas that's been impacted most heavily has been healthcare. The global AI in the healthcare market is anticipated to reach more than $8 billion by 2026 growing at a CAGR of 49.7 percent between 2019-2026.[7] In healthcare, AI is being used in everything from robot-assisted surgeries to clinical trial participant identifiers. AI is revolutionizing the healthcare sector by reducing spending and improving patient outcomes (access, affordability and effectiveness).

> AI in healthcare is transforming the sector by helping reduce expenses while improving patient outcomes

Along with these positive benefits, there are still concerns around AI's utilization in healthcare. As a result, XAI in this industry is built around the pillars of trust, fidelity, transparency, domain sense, generalizability, parsimony and consistency.

How do these pillars work in practice? Let's look at a couple examples.

In the first example, using an AI model to predict heart disease given a patient's records poses a problem of transparency for the clinician. The doctor would want to better understand how the model works for the purpose of improving his or her service. The patient would also want a substantial reason for the prediction made.

To offer another example, domain sense is critical for Emergency Department, or ED, census prediction, which predicts how many patients are in the ED at a particular time. For hospitals, high utilizers of EDs account for a disproportionate number of visits and they are typically for non-emergency conditions, which makes this metric critical.[8] Domain sense is important in ED census prediction because a physician would require a different explanation than someone who plans for staffing. These explanations need to be in the proper language and in the correct context for the key audience.

The need for XAI will only accelerate in the coming years

AI will continue to revolutionize products and services in every imaginable industry: law enforcement, construction, manufacturing and education, just to name a few. Across these industries, AI decisions are carrying heavier ramifications, sometimes to the point of deciding life or death. Think about the AI systems being used in healthcare or those that power driverless cars or the drones that are used in military operations.

The problem is that we currently have little visibility and limited understanding of how AI systems arrive at these decisions. By extension, that means we don't fully grasp how those decisions are applied in the industries where AI systems are heavily used. Machine learning algorithms — especially deep learning approaches — have historically been opaque in terms of our ability to understand their decision-making logic.

Moving forward, this must change. As humans and machines work alongside each other more and more, we must

be able to trust the AI systems that are in place. For that trust to be possible, we must be able to understand the decisions these systems make. Our ability to trust AI suffers in the absence of explainablity. Our expectation is that the machines we work with perform as expected and can explain their reasoning.

With public interest growing in AI and regulations (such as GDPR) demanding a right to explanation from industries that utilize AI systems, companies will have no choice but to update or adopt AI tools that will remove the black box in these algorithms, thereby improving explainability, mitigating bias and improving outcomes for all.

References

1. https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html
2. https://www.gartner.com/en/newsroom/press-releases/2019-01-21-gartner-survey-shows-37-percent-of-organizations-have
3. https://in.reuters.com/article/amazon-com-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idINKC-N1MK0AH
4. https://www.techrepublic.com/article/beware-ais-magical-promises-as-seen-in-ibm-watsons-underwhelming-cancer-play/
5. https://www.juniperresearch.com/press/press-releases/online-fraud-detection-spend-to-hit-9-2bn-by-2020
6. https://www.prnewswire.com/news-releases/global-ai-in-fintech-market-is-forecast-to-exhibit-a-cagr-of-26-21-during-2019-2024--300823214.html
7. https://www.globenewswire.com/news-release/2019/07/08/1879644/0/en/Artificial-Intelligence-in-Healthcare-Market-Size-Worth-US-8-Bn-by-2026.html
8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001776/

Authors

Vijayaraghavan Varadharajan

Principal Research Scientist – Infosys Center for emerging Technology Solutions
Vijayaraghavan_V01@infosys.com

Rohit Chopra

Senior Consultant – Infosys Center for emerging Technology Solutions
Rohit.Chopra@infosys.com

Ramesh N

Principal – Infosys Knowledge Institute
Ramesh_N03@infosys.com

About Infosys Knowledge Institute

The Infosys Knowledge Institute helps industry leaders develop a deeper understanding of business and technology trends through compelling thought leadership. Our researchers and subject matter experts provide a fact base that aids decision making on critical business and technology issues.
To view our research, visit Infosys Knowledge Institute at infosys.com/IKI

For more information, contact askus@infosys.com

Infosys®
Navigate your next