Booz | Allen | Hamilton®

2021 TECHNOLOGY SPOTLIGHT

EXPLAINABLE AI

CONSULTING | ANALYTICS | DIGITAL SOLUTIONS | ENGINEERING | CYBER

Artificial intelligence (AI) enables computers and machines to imitate functions of the human mind such as perceiving, problem-solving, and decision making. However, even if AI decisions are formed utilizing the right analyses, people will not be comfortable with the decisions or trust them if they do not understand how the algorithms and models are working and making judgments. This challenge impacts every sector, from deploying critical government missions to setting policy for social media, which is why monitoring the development of Explainable AI (XAI) will be crucial in the coming years.

XAI is AI that is programmed to describe its purpose, rationale, and decision-making process in a way that can be understood by the average person, and its scope encompasses considerations of bias and ethics. The increased degree of transparency through XAI enables more ethical AI, highlighting the need to minimize discriminatory decisions, respect people's privacy by identifying potential data misuse, and provide people with an understanding of AI decisions in other situations where the potential to cause harm exists.

The federal, fintech, and health management markets are all examples of industry verticals that will be impacted by XAI. In the federal space, billions of dollars have been invested in the Defense Advanced Research Projects Agency (DARPA) XAI program that seeks to develop a suite of explainable machine learning (ML) techniques. Leaders in the ML space use XAI to create more robust information governance with fully auditable decision trails. In the wake of the COVID-19 pandemic, the fintech industry is leveraging XAI so banks can deliver immediate assistance to small businesses and retail customers with rapid loan approval and viable financial products.[1] In healthcare, businesses are automating patient communication and utilizing a suite of deep neural networks for COVID-19 detection and risk stratification through chest radiography to generate clinical recommendations around the clock.[2] Across all of these markets, industry leaders in AI such as Amazon, Google, and IBM are all developing or marketing a component of XAI.

[1] "Temenos Launches New SaaS Propositions to Help Banks Respond to Covid-19 While Accelerating Their Digital Transformation," https://www.businesswire.com/news/home/20200428005713/en/Temenos-Launches-New-SaaS-Propositions-Banks-Respond. *Businesswire.*

[2] "DarwinAI and Red Hat Team Up to Bring COVID-Net Radiography Screening AI to Hospitals, Using Underlying Technology from Boston Children's Hospital," https://www.businesswire.com/news/home/20201116005082/en/DarwinAI-and-Red-Hat-Team-Up-to-Bring-COVID-Net-Radiography-Screening-AI-to-Hospitals-Using-Underlying-Technology-from-Boston-Children%E2%80%99s-Hospital. *Businesswire.*

WHAT IS XAI?

XAI is for people who use ML models to make decisions to establish trust, transparency, and enable oversight of decision making. XAI is an emerging capability that allows for the creation of more transparent AI models. It is a critical component of ethical AI and a mechanism for responsibly using AI. When AI is used in sectors such as medicine, military defense, or data privacy, AI must produce transparent and understandable results that reflect the ethical standards of society. Thus, bias and fairness are characteristics of ethical AI, requiring that AI models' outputs, outcomes, and effects can be understood by those utilizing and impacted by the models.

AI can either learn on its own based on algorithms and utilizing ML, or humans can directly provide inputs into decision making and actively help design AI for explainability. There may be times where this latter learning process is preferrable; for example, if an ML model received all of your health data and indicated you had a life-threatening illness, would you want to know how the model made that determination? Would you also want to ensure that doctors using this AI could verify these claims before determining a treatment regime or going into the operating room? In situations such as these, XAI helps explain the evidence and rationale behind algorithm-led diagnoses and bring the relevant information to the doctor for review. XAI aims to create a suite of ML techniques that produce explainable model results while maintaining a high level of learning performance (or prediction accuracy).

XAI explains decisions for five main purposes:

1. **Analyst Insight** focuses on how machine decisions help augment analyst performance when handling high volume alerts so analysts know where to focus their attention, or when analysts conduct variable relationship discovery where causation and correlation must be determined.
2. **Bias Detection & Bias Mitigation** recognizes that we may not always be able to correct for a bias, but at a minimum, we must be aware of its presence and its potential effects on an algorithm's decision process. Bias mitigation means understanding the impact of unwanted bias and reducing the potential for harm. Bias may be present in both data and algorithms, and both must be investigated throughout the entire AI development process to mitigate potentially harmful bias.
3. **Continuous Improvement** recognizes that human understanding of processes can lead to imperfect decisions. As a result, XAI models clarify the relationships between input and output values to enable continuous optimization and build a symbiotic relationship between the human and the machine.
4. **Quality Assurance** provides consistent, easy-to-read, explainable insights for the interpretation of data and decision outputs of machines (e.g., sanity checks of outputs).
5. **Regulatory Transparency** requires asking questions such as, "*Did the machine do what it was supposed to?*" Transparency capabilities such as the importance rating of decision factors and confidence levels are invaluable to successful XAI—and are also areas where human analyst teams are struggling with documentation today.

Figure 1: AI and XAI Explanation Framework

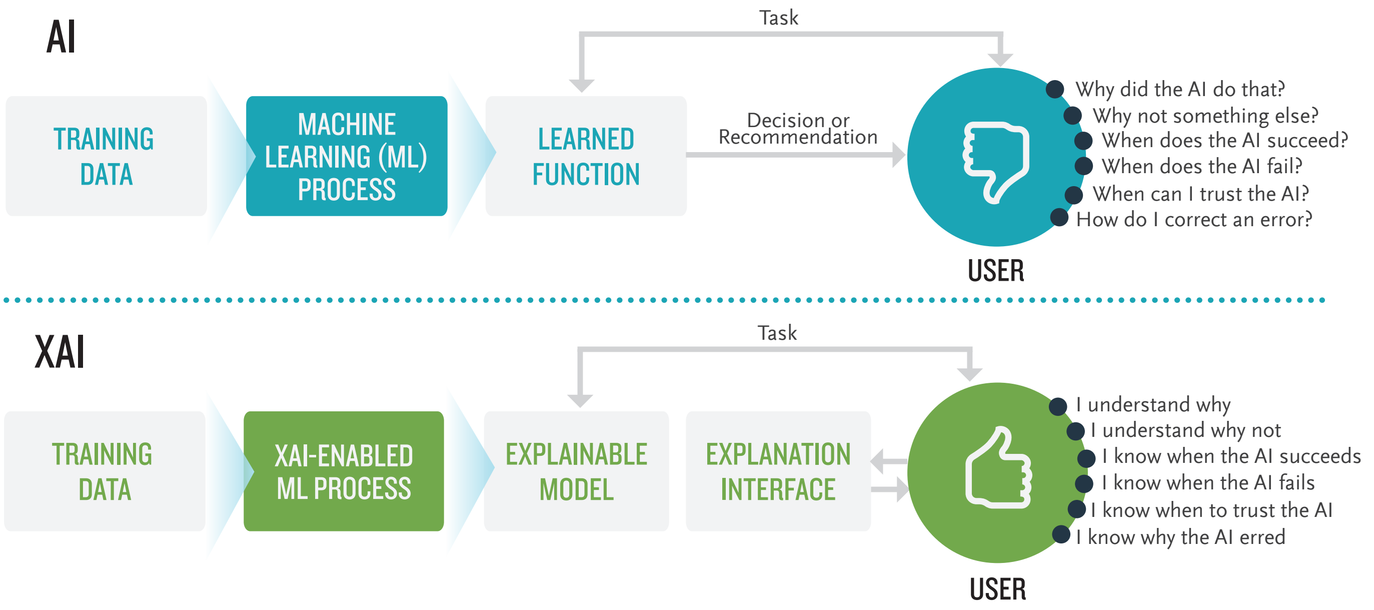


Figure 1 is adapted from a DARPA XAI figure[3] and depicts how an XAI system takes inputs from the current question (i.e., training data) and makes a recommendation, decision, or action while also providing an explanation to the user that justifies its recommendation, decision, or action. The user is now able to decide based on the explanation. Measuring explanation effectiveness depends on the following approaches toward generating explanations:

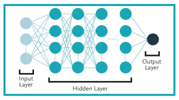
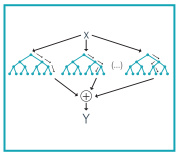
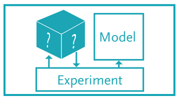
CORRECTABILITY	• Identify errors: Sometimes these errors are difficult to correct, so it is first important to identify them. • Correct errors: Any errors that can be corrected should be, and any errors that cannot be corrected should be made known for human understanding for future usage. • Continuously train and learn: Model or data drift can occur in even the most transparent models, so it is important to continuously monitor the parameters used and the model's understanding of those parameters to ensure the outputs of the model improve and match the expected outcomes.
MENTAL MODEL	• Understand individual decisions: If the decision-making process is intuitive and understandable, then this explanation is effective. • Understand the overall model: If the purpose of a model can be concisely explained, then this explanation is likely effective. • Conduct an assessment to determine the strengths and weaknesses of the model: No model will be perfect, so it is important to understand the limitations of what information the model can accurately provide. • Perform "What will the AI do?" prediction. • Perform "How to intervene?" prediction.
TASK PERFORMANCE	• Determine if the explanation improves the user's decision making or task performance. • Introduce decision tasks to diagnose the user's understanding of the task.

[3] "XAI Figure 2." DARPA. https://www.darpa.mil/ddm_gallery/xai-figure2.png.

TRUST ASSESSMENT	• Perform explainable and appropriate use of data, algorithms, and models to establish trust: Determine that the AI is doing the right thing with the inputs it is given. • Identify bias and potentially harmful outcomes. • Bring awareness to, mitigate, and correct (when possible) for bias in data, algorithms, or models: There will be instances where it is okay to not correct for bias (e.g., where the outcome is not a critical human impact) and other instances when it is not okay to leave bias uncorrected. Either way, there needs to be awareness so that outputs are understood in conjunction with potential bias. • Determine the confidence level of the model's output: Users will often need to act on generated outputs; considering if, when, and how calculations are performed will uncover confidence levels to help the user make a decision and determine the level of trust in the model.
USER SATISFACTION	• Identify the clarity of the explanation (user rating). • Identify the utility of the explanation (user rating): The more uses or applications of the explanation, the more effective it is.

HOW DOES INTERPRETABILITY FIT IN?

Interpretable models are utilized as a learning technique to bring users closer to better models. **Interpretability describes the processes used for understanding the inner workings of ML models. It is a property desirable to ML model developers to understand how algorithms, data, and models work mathematically.** Models are interpretable when humans can readily understand the reasoning behind predictions and decisions made by the model. While interpretability may initially seem like an academic topic more than a practical one, the ability for developers to build mathematically understood models—*answering questions regarding why a model works in some cases and does not in others*—also builds trust in models for non-technical end users.

XAI Models	What It Is	Purpose
VISUAL EXPLANATIONS  	Visually explaining algorithms and data such that people can see the flow and rationale behind an algorithm	Visual explanations are a type of interpretability method that creates scatter plots of visual data or pictures such as heatmaps that can explain the decision of a model. They are tied to concepts similar to saliency maps for computer vision models. Saliency maps plot values that recognize the impact of inputs on a specific decision and the importance of those inputs.
INTERPRETABLE MODELS  	Models that can be assessed by an end user to understand the reasoning for why the prediction provides a recommendation to accept or reject an action	Interpretable models are built to provide an understandable description of the decision logic of AI to an external viewer. Accuracy, efficiency, and explainability are all key factors of interpretable models. Some examples of interpretable models include random forests, decision trees, and linear regressions. These types of models are both transparent and interpretable through natural language explanations, visualizations of learned representations or models, and explanations by example.
MODEL INDUCTION  	Techniques to analyze a model's operations to reverse engineer explanations about the decisions it makes	The goal of model induction is to look at the behavior of the trained system and use that to understand how the model uses data to make decisions. Logistic regression and deep learning are techniques that can be used to build models. Model induction can be performed on models built using these techniques.

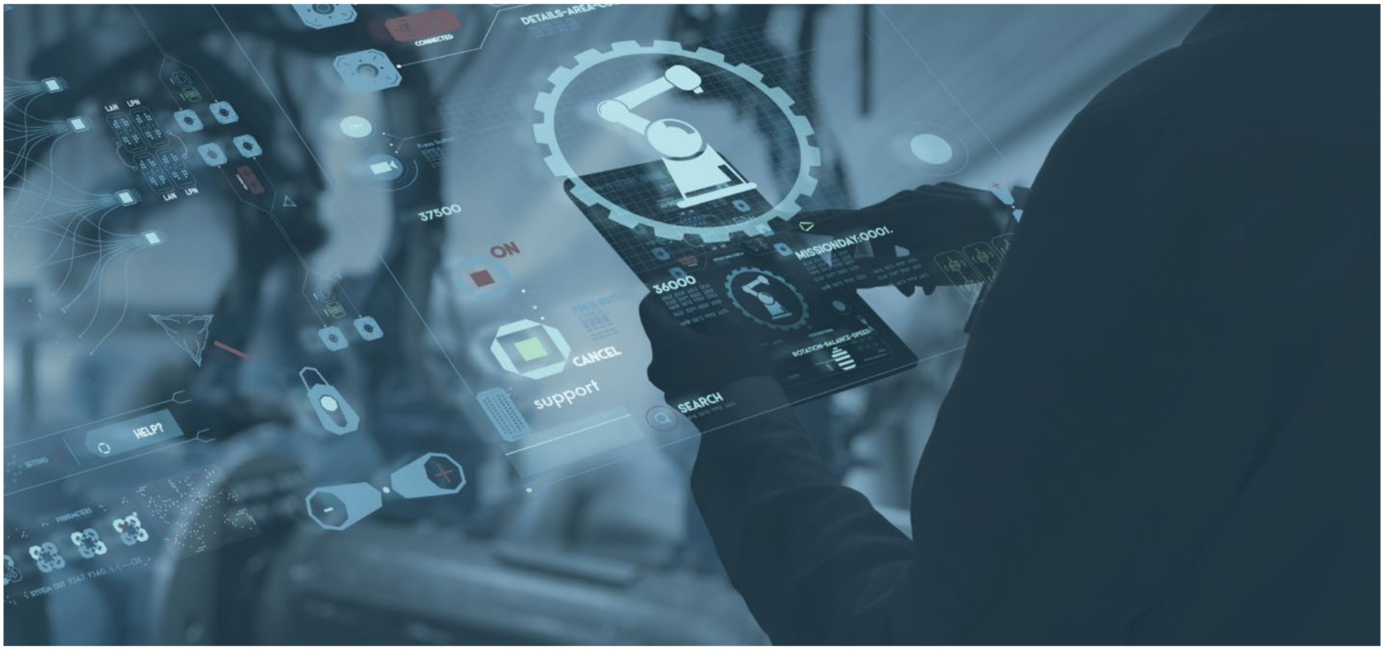
DEEP DIVE: "BLACK BOX" OF AI

Traditional modeling techniques (e.g., regression models and tree-based models) have an understandable relationship between input data and the decisions in the model outputs. These are called white box models. More complex modeling techniques such as deep learning and ensemble models promise increases in model accuracy, but often at the expense of model explainability. These are called black box models because it is much harder to observe and understand the relationship between the model input data and output decisions. Sources of bias, such as unbalanced training data, can impact the reliability of model outputs. Compared to white box models, black box models introduce risk because there is no way to observe what is occurring in a black box model on its own. XAI techniques must be applied to understand the decision-making process.

The better a model is understood, the more justification there is for trust in the model. More trust reduces risk by providing human oversight and understanding of the model decision process. ML models are good at finding patterns in data, and they are also good at optimization: identifying outcomes based on the goals put in place when they are trained. Sometimes these outcomes go beyond what the humans who built the system intended because ML models will often try to find the shortest path to a simple definition of success, regardless of any human values or ethical principles. Interpretability and explainability are necessary for humans to ensure that ML models are making decisions on the correct data, choosing appropriate characteristics on which to base these decisions, and generating outcomes that are equitable, fair, and will not lead to harm.

XAI transforms ML to build trust, clarity, and comprehension through transparency. With XAI, organizations have a direct understanding of an AI's decisions and the factors the produce those decisions. This includes allowing humans to work side-by-side in a human-in-the-loop context, where using dataset collection, dataset annotation, and model validation enable humans to stay informed, via continuous feedback loops, in the decision-making process. In evaluating whether to trust an AI system, people must be involved in each step of the virtuous circle to train, tune, and test a particular algorithm and *must understand the problem-solving process that the AI is undergoing both in real-time and after a decision has been made*. In the next section "XAI Explainability Offerings," **Figure 2** helps outline human-in-the-loop and other common elements.

In summary, XAI gives insights into the data, factors, and decision points used in the suggestion-making process. It is the opportunity to ensure that the decision-making process is transparent. All in all, XAI ought to provide an understanding of what is happening within black boxes and clarify broadly, to the average person, how the decision was made.



XAI EXPLAINABILITY OFFERINGS

Currently, there are many ways to approach XAI and clear techniques available today. Major focus areas for managing XAI include understanding interpretability framework trends and the implementation of standards and policy for XAI. A non-exhaustive list of some typical XAI explainability offerings, examples, and their practical application are captured in **Figure 2**.

Figure 2: Typical XAI Explainability Offerings

EXPLAINABILITY MODELS	**HUMAN IN THE LOOP**	**DETECTION**	**PLATFORM**
The measurement of how well the internal mechanics of a machine or deep learning system can be explained in human terms	The ability to create a continuous feedback loop so humans have an input into the process	The action or process of identifying the presence of something concealed	A centralized location typically on the cloud to analyze the outputs of the algorithms
Example Models: • **SHAP (Shapley Additive exPlanations)** is an XAI approach using game theory and evaluating feature importance to describe the result of an ML model and explainability model. • **PDPs (Partial Dependence Plots)** depict the effect, usually minor, that one or more features have on the effect of an ML model. • **Surrogate Decision Tree** is an interpretable model that can estimate the predictions of a black box model. • **LIME** is an algorithm that can be utilized by ML developers to understand a model by modifying input data and observing how predictions are impacted. • **Integrated Gradients** measures the slope of the model's predicted outcome against the inputted features.	**Examples:** • Regular assessment of AI activities and decisions for human review because model drift occurs and needs real-time alerting of humans for an immediate override in situations when XAI decisions negatively impact human lives • A flowchart of the model's decision process • Comparison in accuracy between models to allow humans to deploy the best model • **What-if scenarios** to stimulate the outcome before running the model (e.g., what-if scenarios that allow data points be manually inputted to measure how robust a model is) • **Heuristic testing:** ML learning from humans • **Reproducibility of AI outcomes:** Quantify the confidence level that work is likely reproducible leveraging both human and machine intelligence (e.g., a researcher "spell check" equivalent). Reproducibility, not repeatability, is the main goal. Computational reproducibility is repeating an experiment using the same data and methods to obtain consistent results.	**Examples:** • **Anomaly Detection:** Detects data drifting and discrepancies and then alerts the user; for example, this can be utilized in an expense reporting system to augment existing rule-based analytics to help address high volumes of expenses reported that trigger as false positives • **Bias Detection:** Identifies sociological biases in data • **Mental Health Detection via AI:** Research from this advancement can be used to determine unsolved human feelings related to affection and closeness. This could then be applied to the fields of psychology to assist individuals with mental afflictions, for example, anxiety or depression.	**Examples:** • XAI platform and ML capabilities embedded with a suite of software through an easy-to-use interface or APIs delivered on-premises, in the cloud, or as a SaaS offering • Accelerates the release of explainable models that will underpin new AI use cases that focus on creating seamless customer journeys and automating manual processes with self-learning capabilities.

ADOPTING XAI WITH ETHICAL AI PRACTICES

IMPLEMENTATION CONSIDERATIONS AND TAKEAWAYS FOR YOUR ORGANIZATION

AI can be used for the common good, but also for adversarial means. In considering the full spectrum of ML, AI, ethical AI, and interpretability, the implementation of XAI will impact capabilities including **data fusion** (the process of receiving data from multiple sources to build more sophisticated models and understand more about a project), **theory-guided data science**, and **adversarial ML** (where minuscule but intentional changes to features can lead to an ML model committing a false prediction). Due to the existence of adversarial ML, it is important to stay competitive by driving decisions on which AI algorithms can be used and ensuring the confidentiality of these algorithms.

Every notable AI company is developing or marketing a component of XAI; however, there are very few companies that focus solely on providing XAI models. While the XAI startup market is still very small with fewer than 50 companies identified, they are also developing XAI as an additional feature of their pre-existing products and platforms. Taking all of that into consideration, it is important to continue to monitor the XAI market because it is a growing field and changes occur rapidly.

Leaders looking to explore XAI and XAI solutions should consider the following recommendations:

High-Impact Industries: XAI ensures that data is ethically used by organizations in an auditable and logical way to defend the fairness of decisions. Recognize that XAI raises the standard for most critical public-facing technologies across industries that operate in fields reliant on sound ethics. Also, proposals for XAI ethics and law are quite amorphous, so develop AI regulations with explainability in parallel with technology.

Partnership Prioritization: XAI will enable federal leaders to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners. Partner with firms working in spaces where investment decisions rely heavily on AI to lighten their compliance load. XAI platforms with enhanced data insights will automate workflows.

Technical Investments: Discover how processes in place to query automated decisions can build trust between the public and autonomous systems to foster wider adoption. Revolutionize the automated black box process of compliance outreach to investors to collect outstanding information through XAI. Collaborate with companies tracking the growing role of XAI that put data into context when a business user needs to act as part of the business workflow.

Workforce Transformation: Humans will need to continually tweak performance to trust the AI system is working correctly. Encourage the development of XAI to grow this area. Join companies that see the value in building flexible and interpretable models that can work in collaboration with experts and their domain knowledge.

Booz Allen recognizes that **ethical AI is not just a buzzword, but about human impact driven by values, and it requires the right tools and right governance**. It is about continuing to keep AI rooted in deep scientific and technological rigor by incorporating built-in mechanisms for tracking data, model accountability, and auditability. The pathways that empower us to design balanced AI are awareness, access and redress, accountability, explanation, and data provenance.

Our AI guiding principles align with Booz Allen's ethics and maintain a deeper connection to and awareness of the societal and environmental impacts. At the same time, we ensure that humans continue to be at the core of every AI system retaining ultimate control—calling out any unintended consequences and potential dual uses of technology through bad or neutral actors while building consumer trust. Consumers need to believe that all factors will be considered to ensure the solution works correctly for all types of people. One way we do that is by building AI teams that are meaningfully diverse and inclusive to have a wide range of perspectives and experiences to draw on. Booz Allen understands that the future of XAI and ethical AI will depend on humans holding each other accountable to build a better tomorrow.

FOR MORE INFORMATION:

The content in this 2021 Technology Spotlight Series originates from Booz Allen's emerging technologies capabilities as part of our innovation agenda, and this spotlight on explainable AI and ethical AI was developed by Booz Allen's Technology Scouting team. For more in-depth AI guidance, please visit **BoozAllen.com/ai**.

About Emerging Technologies at Booz Allen

For over a century, Booz Allen has been at the forefront of technology and strategy. We provide strategic advisory services, design, build, and deploy solutions across sectors to a wide range of federal government organizations. We are the only management and technology consulting firm that has invested to establish an innovation network exclusively focused on scouting dual-use technologies for federal government missions. As a result, we successfully partner with the tech community to showcase a wide range of capabilities and deliver innovative solutions to our clients.

Our deep pool of technical talent, access to critical networks in innovation hubs, and experience performing tech scouting across technologies and sectors uniquely position Booz Allen to serve clients who seek to be at the cutting edge.

The content in this 2021 Technology Spotlight Series originates from our emerging technologies capabilities and aligns with our innovation agenda.

To learn more, visit BoozAllen.com (NYSE: BAH)