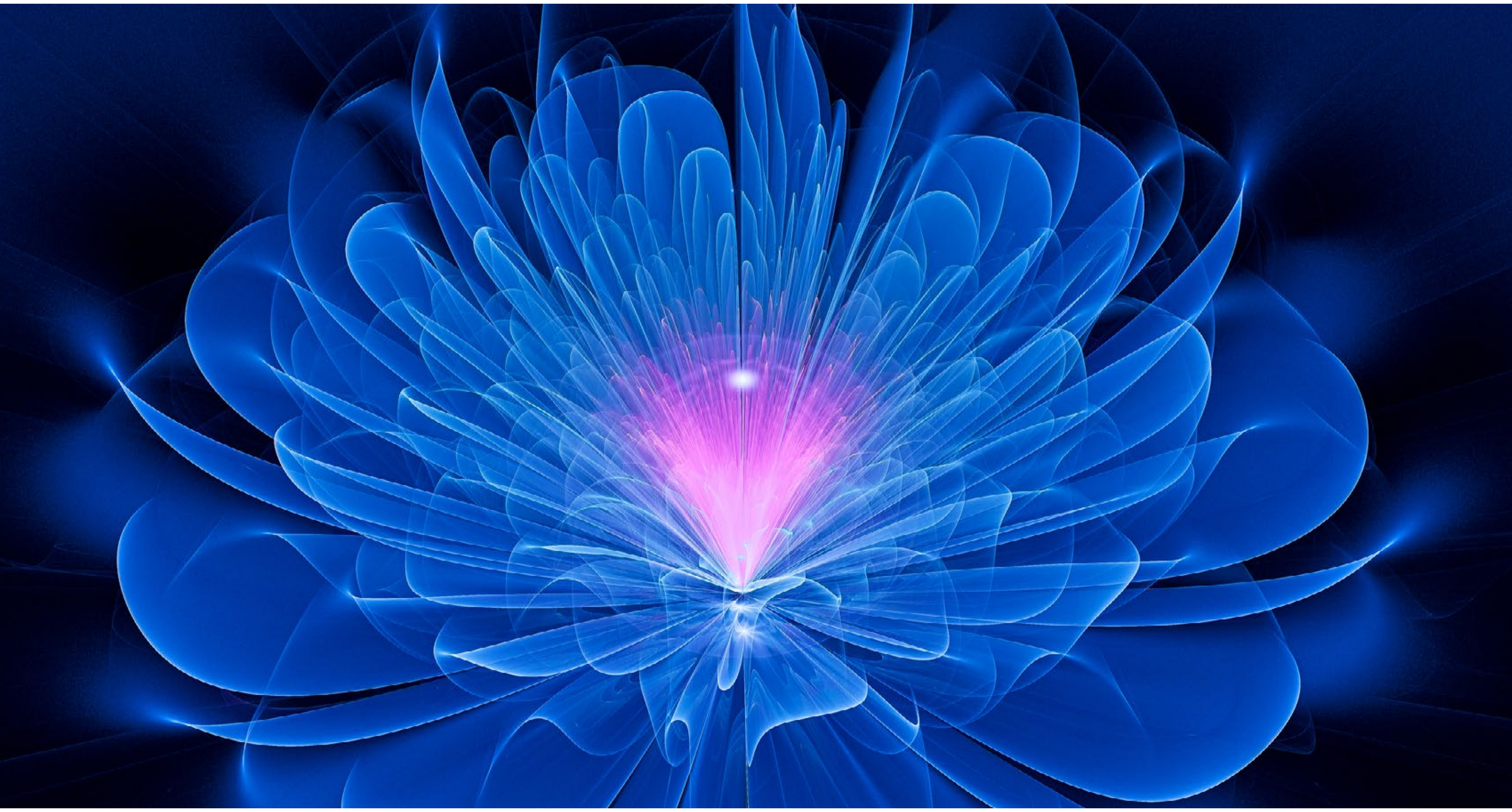
QuantumBlack
AI by McKinsey

Building AI trust: The key role of explainability

AI systems are powerful but often operate like “black boxes,” shrouded in mystery. Here’s how companies can shed some light and drive adoption of AI solutions that users trust and understand.

by Carlo Giovine and Roger Roberts
with Mara Pometti and Medha Bankhwal



November 2024

Artificial intelligence has the potential to deliver massive gains in economic productivity and enable positive social change around the world. So it's little surprise that the number of companies adopting AI-powered software, tools, and platforms, including generative AI (gen AI), has surged throughout 2024. But that enthusiasm has been accompanied by a fair amount of trepidation: in McKinsey research, 91 percent of respondents doubt their organizations are "very prepared" to implement and scale the technology safely and responsibly.[1] Such doubt is understandable. Along with its potential to boost productivity and innovation, gen AI in particular poses novel risks—for example, hallucinations and inaccurate or biased outputs—which threaten to undermine trust in the technology.

To capture the full potential value of AI, organizations need to build trust. Trust, in fact, is the foundation for adoption of AI-powered products and services. After all, if customers or employees lack trust in the outputs of AI systems, they won't use them. Trust in AI comes via understanding the outputs of AI-powered software and how—at least at a high level—they are created. Organizations increasingly recognize this. In a McKinsey survey of the state of AI in 2024, 40 percent of respondents identified explainability as a key risk in adopting gen AI. Yet at the same time, only 17 percent said they were currently working to mitigate it.[2]

This conundrum has raised the need for enhanced AI explainability (XAI)—an emerging approach to building AI systems designed to help organizations understand the inner workings of those systems and monitor the objectivity and accuracy of their outputs. By shedding some light on the complexity of so-called black-box AI algorithms, XAI can increase trust and engagement among those who use AI tools. This is an essential step as AI initiatives make the difficult journey from early use case deployments to scaled, enterprise-wide adoption.

Why invest in this capability: Getting ROI from XAI

As with any investment in an uncertain environment, organizations seeking to enhance AI explainability must consider the benefits and costs to decide how and when to act in the absence of perfect information on the potential upside and risks involved. Today's AI landscape is fraught with uncertainty, and in this context, leading AI labs like Anthropic are making bets that investments in XAI will pay off as a path to differentiation in a crowded field of foundation model builders (see sidebar "The evolution of XAI and today's challenges"). Meanwhile, enterprises are seeking to meet the expectations of their stakeholders and regulators.

One thing is certain: demand for XAI is rising. As global AI regulations begin to take shape, the need for explainability and interpretation is increasing, with more organizations seeking guidelines on how to determine what level of explainability to adopt and how much information to release about their models. The EU AI Act, for example, imposes specific transparency requirements for different AI use cases classified according to its risk-based framework. For example, in the case of high-risk AI systems—such as systems used in recruitment, like résumé-ranking software—organizations are required to provide information about the system's capabilities, limitations, data lineage, and the logic behind the decisions it makes.

Imagine driving a car. Setting a speed limit of 45 miles per hour is useless if your vehicle lacks a speedometer that indicates where you are relative to the standard. Similarly, to respond to AI regulations, organizations need methods that provide visibility into how AI models are built and how they can be tested before release. Organizations also need observability solutions that provide sufficient insight into their AI models to ensure they comply with regulations and their own values and standards. This raises crucial questions: Are organizations prepared to deliver this level of transparency? Do they have the necessary capabilities and technologies in place? Have

[1] "Implementing generative AI with speed and safety," *McKinsey Quarterly*, March 13, 2024.
[2] "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," McKinsey, May 30, 2024.

The evolution of XAI and today's challenges

The field of AI explainability has evolved significantly in recent years. Early AI tools, employing rule-based systems and decision trees, were relatively simple and transparent by design. However, as machine learning models have grown more complex, it has become more difficult to trace the reasons underpinning their decision-making processes. The early 2000s saw the development of methods like local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP), which provided insights into individual predictions of complex models. Google introduced its What-If Tool, enhancing model transparency through interactive visualizations; IBM released the AI Explainability 360 tool kit; and DARPA produced an Explainable AI (XAI) program, which further advanced the field by developing comprehensive tool kits and techniques to interpret AI models.

In the meantime, several highly publicized missteps have highlighted the growing need for AI explainability. In 2016, for example, a ProPublica investigation into the COMPAS algorithm, used by US courts to assess the likelihood of a defendant reoffending, revealed systematic bias against African American defendants. Unfortunately, addressing these concerns is no simple matter. One major issue is the increasing complexity of advanced large language models (LLMs), which rely on deep neural networks and often operate as black boxes, with opaque decision-making processes. And the lack of access to the architecture of proprietary models makes it difficult to understand how they operate. Previously, teams could control fairness by curating training data and applying calibration techniques. However, today's LLMs resist such control, making explainability difficult. Finally, organizations increasingly face a trade-off between model accuracy and interpretability: more complex models and LLMs often achieve higher accuracy but at the cost of being less interpretable and harder to explain.

Innovators are focused on these issues, and some strides have been made. Anthropic, for example, has provided significant improvements to techniques for LLM explainability and interpretability. Tools to interpret the behavior of language models, including OpenAI's transformer debugger, are new and only beginning to be understood and implemented. An example of how tech companies are incorporating explainability tools into their platforms is Google's Vertex Explainable AI, which enhances understanding of generative AI and LLMs through feature-based and example-based explanations that give users insights into model predictions by identifying influential features in complex generative models like transformer-based LLMs. In addition, recent community-driven research, like work on behavior analysis at the head level of LLM architectures, reflects growing momentum toward unpacking model behaviors. The scale and complexity of more mature techniques for unpacking these intricate systems present unprecedented challenges, but even if much work remains, we anticipate progress in the coming years.

platforms and innovators created reliable methods of measurement?

XAI is best thought of as a set of tools and practices designed to help humans understand why an AI model makes a certain prediction or generates a specific piece of content. The ultimate goal is to ensure that these outputs are of high quality, untainted by bias, inaccuracy, or hallucination. This requires several kinds of investment—in tools, people, and processes. Building more explainable AI and machine learning solutions requires deployment of new technology in the software delivery life cycle (SDLC) from the start, when models are trained and tested or as pretrained models are fine-tuned, ending when code moves into production and ongoing monitoring and observability are needed. Expertise in XAI techniques must be built via hiring and/or training, and the experts must be integrated into the SDLC right from the conception of new AI-powered offerings. These experts can form an XAI center of excellence (COE) to provide expertise and training across teams, reshaping the software development life cycle and assuring coordinated enterprise-wide investments in tools and training. The COE also can address the need for additional compute power and cloud consumption to deliver the additional training, post-training, and production monitoring essential to enhancing explainability.

How can we ensure a return on these investments in technologies that are often in their early stages? While XAI is still emerging from research-focused efforts in academia and R&D labs into real-world applications, its benefits are far more tangible than commonly thought. We see five areas where XAI can deliver a return that drives positive ROI:

1. *Operational-risk mitigation.* By revealing how AI models process data and produce results, XAI enables early identification and mitigation of potential issues, such as bias or inaccuracy, reducing the risk of operational failures and reputational damage. For example, many financial services companies use AI in fraud detection yet often struggle to control or understand why their AI systems behave the way they do—a potential problem, considering how damaging false positives for fraud can be to both the company and the customer. Explainability can increase organizations' understanding of why models flag certain transactions, allowing them to fine-tune their systems or introduce greater human oversight.

2. *Regulatory compliance and safety.* XAI ensures that AI systems operate within industry, ethical, and regulatory frameworks, minimizing the risk of noncompliance penalties and protecting brand integrity. In human resources, for example, many recruiters use AI tools to help screen and select candidates. Explainability ensures that hiring decisions are fair and based on relevant criteria, avoiding bias and discrimination.

3. *Continuous improvement.* XAI supports the ongoing refinement of AI systems by providing insight into the way the systems function, fostering targeted debugging and iterative improvements that help developers keep AI systems aligned with user and business expectations. Many online retailers, for example, use explainability to improve their recommendation engines so they better match recommendations with customer preferences.

4. *Stakeholder confidence in AI.* By attempting to make AI systems understandable, XAI shifts the focus from the technical functioning of models to the users of those models, fostering a human-centric approach that empowers users by boosting their understanding of how AI outputs are generated. In the healthcare sector, for example, AI systems increasingly are used to identify potential illness. XAI can help doctors better understand why these models behave the way they do, driving confidence and adoption.

5. *User adoption.* XAI helps organizations monitor the alignment between a model's outputs and users' expectations. Greater alignment, in turn, increases adoption, satisfaction, and ultimately top-line growth through innovation and change management.

XAI is not and cannot be an afterthought. Rather, by integrating explainability into the design, development, and governance of AI systems, organizations can unlock tangible value by facilitating adoption, improving AI model performance, and boosting user confidence. Nor is XAI simply a compliance issue or requirement. It's a strategic enabler of adoption, trust, and ultimately business success—a crucial tool for maximizing the value of AI technologies across the organization.

XAI is a catalyst for a human-centered approach to AI

As organizations consider investing to capture a return from XAI, they first must understand the diverse needs of the different constituencies involved and align their explainability efforts to those needs. Varied stakeholders, situations, and consequences call for different types of explanations and formats. For instance, the level of explainability required for an AI-driven loan approval system differs from what is needed to understand how an autonomous vehicle stops at an intersection. A high-risk scenario, such as a cancer diagnosis, could demand a precise explanation provided rapidly, while the rationale for a restaurant recommendation can be handled with less urgency. Whatever the circumstances, the type of explanation needed, which informs the XAI technique required, should be derived with a human-centered approach—one rooted in the needs of the people seeking explanations of an AI's outputs (see sidebar "A human-centered approach to AI explainability").

It's helpful to think of AI explainability as a bridge across a chasm. On one side are the engineers and researchers who study and design explainability techniques in academia and research labs, while on the other side are the end users, who may lack technical skills but still require AI understanding. In the middle, bridging two extremes, are AI-savvy humanists, who seek to translate AI explanations

A human-centered approach to AI explainability

One of the advantages of XAI is that it places humans at the center of AI efforts. “Data storytelling plays a crucial role in bridging the gap between human understanding and AI,” says Giorgia Lupi, a partner at the design firm Pentagram and creator of the Data-Humanism Manifesto. “By translating the machine’s thought processes into narratives that resonate with our natural ways of learning, we can make AI’s complex logic more accessible—and therefore truly useful.”

According to Lupi, explainability efforts are fundamentally about humanizing the machine’s inner workings and framing AI’s data as stories that reveal its logic. “When we embrace storytelling to articulate how AI ‘thinks,’ we invite people to connect with this new kind of teammate, fostering understanding through a shared language of curiosity and exploration,” she says. “It’s not just about demystifying AI; it’s about finding poetry in how we learn to work alongside it. Think of AI like another member of the team that is from another culture. We, as humans, need to be open to learning how AI reasons, even if it’s not immediately intuitive for us.”

developed by researchers and engineers to respond to the needs and questions of a diverse group of stakeholders and users. This emerging talent will be the key to designing XAI that works for all.

Stakeholder needs can be broken down into six personas, each benefitting from different techniques and explanations:

— *Executive decision makers* require enough understanding and information about models to be accountable for their actions with respect to customers and employees—specifically, to ensure that models behave in alignment with the organization’s strategies, brand ethos, and values.

— *AI governance leaders* constitute a cross-functional group—drawn from functions like legal, risk, information security, engineering, and product—that is responsible for shaping AI systems in accordance with policies, standards, and regulations.

— *Affected users* need explanations about the outcomes they get from AI models.

— *Business users* require insights to enhance everyday decision making, improve processes, and optimize operational efficiency.

— *Regulators/auditors* require explanations and interpretability from models to make sure they are safe and compliant as rules and regulations evolve.

— *Developers* need explanations of the models’ functioning so they can improve and debug these nondeterministic systems, add post-training enhancements, and ensure the AI models deliver expected outcomes.

Beyond these different stakeholders, varying contexts and risk scenarios influence the format of the explanations provided. Explanations can take the form of data visualizations or text reports and will vary in technical detail. Understanding the specific needs of each stakeholder at a particular time is essential to providing effective and meaningful AI explanations that meet their unique needs.

How does XAI work, and what techniques are available?

To meet these diverse needs, the XAI community continues to create new explainability techniques, which involve algorithms to make the decision-making processes of AI models more transparent. These can be grouped based on stakeholders’ intents and goals. In general, techniques can be categorized along two dimensions: when the explanation is produced (before or after the model is trained) and the scope of the explanation (global or local).

The first macro category of XAI techniques comprises “post-hoc methods,” which involve analyzing models after they have been trained, in contrast to “ante-hoc methods,” which refer to intrinsically explainable models, like decision trees. For example, when an HR department seeks to predict which employees may be more likely to leave an organization, a decision tree can transparently show why certain employees are identified as turnover risks based on factors like job satisfaction, tenure, and workload. In this case, an ante-hoc explanation is inherent in the AI model and its functioning. By contrast, an AI model that uses neural networks to predict the risk of a condition like diabetes or heart disease in a healthcare setting would need to provide an explanation post hoc, or after the results are generated. Most often, it does this by applying techniques (like SHAP or LIME) to identify which factors (for example, age, lifestyle, or genetics) contribute most to the risk score and determine whether the risk score is accurate and unbiased.

The second dimension differentiates between global and local explanations. Global explanations help us understand how an AI model makes decisions across all cases. Imagine a bank that uses an AI model to assess loan applications. By using a global explanation tool (such as Boolean rule column generation), the bank can see which factors—such as income, debt, and credit score—generally influence its loan approval decisions across all customer segments. The global view reveals patterns or rules that the model follows across the entire customer base, allowing the bank to confirm that the model aligns with fair-lending regulations and treats all customers equitably. Deploying XAI algorithms atop a loan application model provides loan officers with a rich base of information and statistical insights to understand the factors driving the system's decisions, allowing them to confidently explain approval patterns to customers and regulators.

Local explanations, in contrast, focus on specific decisions. Consider a healthcare setting, in which a doctor uses AI to help diagnose patients. By applying a local explanation tool (such as SHAP), the doctor can see precisely why the model predicted a certain condition for that specific patient—showing, for instance, that a patient's age, medical history, and recent test results influenced the model's prediction. This level of detail can help doctors understand the model's reasoning for individual cases, so they have more trust in its recommendations and can provide more informed, personalized care.

Beyond these two macro categories, AI explainability techniques also can be mapped to the needs of different personas according to the use cases and their context—say, to help developers debug systems to boost accuracy or strengthen bias detection or assist product leaders in improving personalization efforts. Much academic and research work in AI labs is ongoing to enhance and improve the range of capabilities available to meet the rising demand for XAI and, when paired effectively with user-centered design, to meet the needs of the six personas described earlier in this article.

How to start with XAI

Given that the appropriate techniques used to get explanations on AI models are informed by the personas that need explanations in different contexts, organizations should consider several steps for embedding explainability methods into their AI development.

Build the right XAI team

Organizations should create truly cross-functional teams, comprising data scientists, AI engineers, domain experts, compliance leaders, regulatory experts, and user experience (UX) designers. This diverse group ensures that the explainability efforts address technical, legal, and user-centric questions. Data scientists and AI engineers will focus on the technical aspects, while domain experts and designers provide context-specific insights and shape the content and format of the explanations.

Establish the right mindset

The XAI team should consist of builders, not judges. It should focus on accelerating innovation while assuring the right insights are wrapped around the products or services being built. To do this, the team needs to engage while ideas are being shaped into buildable concepts, not at some later stage. Early involvement helps establish a human-centered engineering culture in AI while avoiding downstream conflicts between "engineers" and "explainers."

Define clear objectives

Set clear goals for AI explainability for each stakeholder persona. Determine what needs to be explained, to whom, and why. This involves interviewing key stakeholders and end users and understanding their specific needs. Establishing clear goals helps with selecting the right techniques and tools and integrating them into a build plan.

Develop an action plan

Create a strategy to embed explainability practices, from the design of AI solutions to the way explanations will be communicated to different stakeholders. The former ensures the adoption of explainability tools across the entire AI life cycle. The latter involves deciding on the format (visualizations, textual descriptions, interactive dashboards) and level of technical detail (high-level summaries for executives versus detailed technical reports for developers). Ensure that the explanations are clear, concise, and tailored to the audience's understanding.

Measure metrics and benchmarks

AI explainability also demands a strong push for industry-wide transparency and standardized benchmarks that not only help users understand AI systems better but also align with regulatory expectations. For instance, Hugging Face's benchmarking efforts, in which it measures and tracks compliance with the EU AI Act, and the COMPL-AI initiative's focus on assessing and measuring model transparency are important steps toward greater accountability. As these frameworks mature, they will be crucial for fostering trust and advancing responsible AI practices across the industry.

Select or build appropriate tools

Adopt and integrate explainability tools that align with the organization's needs and technical stack. Some widely used tools include open-source algorithms such as LIME, SHAP, IBM's AI Explainability 360 tool kit, Google's What-If Tool, and Microsoft's InterpretM. Ensure that the XAI core team keeps an eye on the rapid innovation in this domain.

Monitor and iterate

Continuously monitor the effectiveness of the explainability efforts and gather feedback from stakeholders. Use this feedback to iterate and improve the explainability processes. Regularly update the models and explanations to reflect changes in the data and business environment.

By following this path, organizations can successfully embed explainability into their AI development practices. Then AI explainability will not only enhance transparency and trust but also ensure that AI systems are aligned with ethical standards and regulatory requirements and deliver the levels of adoption that create real outcomes and value.

As enterprises increasingly rely on AI-driven decision making, the need for transparency and understanding becomes paramount across all levels of the organization. Those that fail to build trust will miss the opportunity to deliver on AI's full potential for their customers and employees and will fall behind their competitors.

Ultimately, trust will be a key to responsible adoption of artificial intelligence and bridging the gap between a transformative technology and its human users. However, trust cannot stand alone. As a bridge, it must be supported by strong pillars. For AI trust, those pillars are explainability, governance, information security, and human-centricity. Together, these pillars will enable AI and its human users to interact harmoniously, making AI work for people and not the other way around—and providing a foundation to ensure that AI systems deliver tangible value to users while preserving respect for human autonomy and dignity.

Carlo Giovine is a partner in McKinsey's London office, where **Mara Pometti** is a consultant; **Roger Roberts** is a partner in the Bay Area office, where **Medha Bankhwal** is an associate partner.

This article was edited by Larry Kanter, a senior editor in the New York office.

Designed by McKinsey Global Publishing