Opening the Black Box

The value of explainable AI in financial services
Artificial intelligence is changing the world around us.\(^1\) Every day, AI systems are buying and selling millions of financial instruments, assessing insurance claims, assigning credit scores and optimizing investment portfolios. As applications grow, it is not enough for AI systems to perform well. We need to understand how they work so we can trust them enough to use them to their full potential.\(^2,3\)

The challenge for modern AI, unlike previous technologies, is that how and why it works isn’t always obvious, even to the technology’s creators. Many of the advanced machine learning algorithms that power AI systems are inspired by the human brain, yet they lack the human ability to explain their actions or reasoning.

Thankfully, there’s an entire research field working towards describing the rationale behind AI decision-making: Explainable AI (XAI). Momentum in the field is growing as AI systems demonstrate performance and capabilities far beyond previous technologies, but encounter hurdles of practicality and legal compliance. For companies putting AI to work, XAI will be a key factor in successful implementations.

---

\(^1\) AI, specifically deep learning, is beating records in image recognition, speech recognition, predicting the activity of potential drug molecules, analyzing particle accelerator data and reconstructing brain circuits. Furthermore, AI has produced extremely promising results for natural language understanding, particularly for the tasks of topic classification, sentiment analysis, question answering and machine translation.


\(^3\) Gerald Fahner, Developing Transparent Credit Risk Scorecards More Effectively: An Explainable Artificial Intelligence Approach, Data Analytics, 2018. Available at: https://www.thinkmind.org/index.php?view=article&articleid=data_analytics_2018_1_30_60077
Explainable AI in Financial Services

Explainability techniques will prove to be especially valuable in financial services, where the low signal-to-noise ratio typical of financial data demands a strong feedback loop between user and machine.4 AI solutions that do not leave room for human feedback to guide outputs risk never being adopted in favor of traditional approaches that rely on domain expertise and experience honed over many years. Regulation, too, raises the stakes by preventing AI-powered products from even entering the market if they are not auditable.

Market forecasting and investment management

Time series forecasting methods have grown in prominence across financial services. They are useful for predicting asset returns, econometric data, market volatility and bid-ask spreads, to name a few. But their success is limited due to their dependence on historical values. Because they can lack disparate meaningful information of the day, using time series to predict

4 Lael Brainard, What Are We Learning about Artificial Intelligence in Financial Services?, Speech at At Fintech and the New Financial Landscape, 2019. Available at: https://www.federalreserve.gov/newsevents/speech/brainard20181113a.htm
the most likely value of a stock or market volatility is very challenging. Complementing these models with explainability methods could allow users to understand the key signals the model uses in its prediction, and interpret the output based on their own complementary view of the market. This would in turn enable a synergy between the domain expertise of finance specialists and the big data crunching abilities of modern AI.

Explainability techniques enable similar human-in-the-loop AI solutions for selecting a portfolio. An investor might not choose to pick the suggested portfolio with the highest reward if the risk associated with it seems too large. On the other hand, a system that also provides a detailed explanation of the risks, e.g., how they are uncorrelated with the market, would be a powerful investment planning tool.

Credit scoring

Assigning or denying credit is a consequential decision that is well regulated to ensure fairness. Many opportunities for AI applications in credit scoring are dependent on the ability of an AI application to provide a robust explanation of its recommendation. Beyond compliance, the value of XAI can be seen for both the client and financial institution: clients can receive explanations that give them the information they need to improve their credit profile, while service providers can better understand predicted client churn and adapt their services. XAI in credit scoring can also help with derisking: for instance, an XAI model might provide an explanation of why a pool of assets has the best distribution to minimize the risk of a covered bond.

As we explore in this discussion paper, successfully applying explainability will mean taking a user-centric approach to XAI that starts at the beginning of the solution development.

---

The Business Value of Explainable AI

XAI has the potential to encourage AI adoption and enable trust in AI, expand the range of possibilities for AI applications in regulated industries, speed up deployment and the debugging of AI solutions and detect bias and monitor AI outcomes.

Explainability effects

Business Value

- Detecting Bias
- Regulatory Compliance
- Establishing Trust
- Enhanced Debugging
- Improved Robustness

Shared Goals

- Ethical AI
- Mass Adoption
- Robust AI

Put simply: XAI encourages trust, trust drives adoption, and adoption drives business value.

AI has shown great promise to transform business, but it faces the same challenge of every previous technology: adoption. To drive AI adoption, explainability is key. ⁶

There is always a transition period in which a user community validates the capabilities and limitations of any technology. This transition is a critical moment of building trust between users and their tool. In the case of AI, that trust hinges on whether the user can trust the decisions presented to them by the machine. Explainable AI will go a long way towards building credibility for an AI technology if we know both why and how the system is reaching its decisions.

These explanations can transform the solution they are integrated into, not only creating more value for the end-user, but also can help contribute to other business needs like regulatory compliance, detecting bias, protections against...

The Future of Compliance

Data-rich, regulated areas such as financial services offer some of the most promising near-term applications for AI decision-making. Though, in certain circumstances, individuals have a right to an explanation of decisions that impact them. There is immense value to potentially be unlocked by integrating XAI capabilities into an AI system assisting with those decisions.

Governments and regulators are beginning to study AI and its impacts, and explainability could play a part in future compliance with emerging regulation.\(^7\) Lawmakers in the United States have introduced the Algorithmic Accountability Act, which calls on large companies to check their algorithms for bias. EU politicians, who have led the way on privacy legislation with the GDPR, have discussed similar accountability measures in the trade bloc.\(^8\) In Singapore, the FEAT principles are setting up the new rules for technology-driven finance.\(^9\) Pursuing XAI now could open up the opportunity for a wider number of AI applications in regulated industries and beyond in the future.

adversarial techniques — those seeking to game AI systems by feeding them specially tailored data — and speed of deployment.

XAI could also allow for organizations to ensure the fair treatment of protected classes, such as sex and national origin. If a protected class is found to have a high significance for model predictions, that model could be said to be biased.

These are not hypotheticals: controversies have already arisen around the AI systems used to assign credit scores and give out bank loans, which could end up incorrectly denying loans to those from minority communities.\(^10\) In this case, the AI systems were making decisions based on features within the data that would or should be irrelevant to a person facing the same choice.

Lastly, XAI can also deliver business value through speed. As explainability methods mature and are integrated into processes to build and deploy AI systems, they can shorten development times (and time to market) by making it easier to debug, enhance and tune the models.

The potential for XAI becomes ever more important as AI adoption becomes more widespread. At nearly every stage of AI implementation, XAI could become a competitive differentiator for those who integrate it into their AI systems.


\(^9\) Monetary Authority of Singapore (MAS), Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector, mas.gov.sg, 2019. Available at: https://www.mas.gov.sg/publications/monographs-or-information-paper/2019/FEAT

XAI is the effort to provide the reasoning behind an AI model, both its output and its inner workings.\textsuperscript{11,12} Some of the algorithms that power modern learning systems, including traditional machine learning models and modern neural networks, are built on a complex mathematical foundation. Because of that complexity, many current machine learning models are black boxes — you put something in, you get something out, and it’s not interpretable what happened in the middle.

While AI models have gotten much better at certain tasks, they still do not reason like we do. When we talk about AI performance versus a human benchmark, it’s about the input and the output — not what happens in between to generate the output. Explainable AI is about figuring out the activity happening in those black boxes: how data is being processed by the model, and how that affects the model’s output.

One easy example to aid in understanding the need of XAI is in image recognition, where modern AI made its first big gains back in 2011 and 2012.\textsuperscript{13} To the right are several images unrecognizable to a human that were misclassified by state-of-the-art AI algorithms.\textsuperscript{14}

It’s easy to see how the algorithm in the examples to the right was tripped up by certain images: the stitching on what was inaccurately labeled a baseball, or the coloured buttons on what it labelled a remote control. Though the AI algorithm is able to recognize the common features of those objects, it fundamentally does not perceive the world in the way humans do. We need XAI to help us understand how AI is actually “seeing” the world so that we can fix it or work productively with it.

A popular explainability technique known as “feature attribution” can be a resolution to image classifiers making more subtle mistakes. Using feature attribution, it’s possible to demonstrate which pixels or sections of an image are most important to the given output. This kind of explanation makes it easier to spot the spurious correlations like in the previous images.

\textsuperscript{14} Anh Nguyen, Jason Yosinski, and Jeff Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 427-436), 2015. Available at: https://arxiv.org/abs/1412.1897
In one oft-referenced example, an algorithm had mislabeled an image of a Siberian Husky as a wolf. Researchers were able to extract the reasoning for the decision. It wasn’t the lupine qualities of the dog itself, such as the fur or the snout or the ears, but that the Husky, like the wolves in the photos the algorithm had already seen, was surrounded by snow.

We point to image classification examples because they make for easy illustrations of why it’s useful to have an explanation of what a model deems important. Much of the field focuses on model explanations like the one to the left, and researchers are currently working on transposing those techniques developed on image- and text-based applications to build explainable and interpretable time series forecasting models that are applicable in financial services.\textsuperscript{16,17}

Researchers discovered that this Husky was incorrectly misclassified as a wolf because the algorithm was looking at the snowy background and not the dog.

\textbf{Technical Deep Dive}

Shapley additive explanations (SHAP)\textsuperscript{18} are a feature attribution technique that have been growing in popularity for explaining credit scoring models. This article from Zest AI is an excellent description of why it is critical to understand the business context and user perspective. Otherwise, the ways that data scientists use techniques like SHAP may not meet regulators’ expectations or industry standards for reason codes or other types of explanations that are typically provided alongside non-AI models.

Yet such feature attribution techniques are often insufficient on their own in financial services applications because they only show the relationship between the input and the output. They provide little information about the inner workings of a model, which can miss the cause of a bad output (see Technical Deep Dive). These “post-hoc explanations” are called “post-modelling explainability”, but explainability techniques can also be applied before (“pre-modelling explainability”) as well as during (“explainable modelling”) the modelling stage.

In addition to where the explanation needs to happen, the relevant question to answer also varies based on the asker: a developer creating an AI model might prefer robust explanations that describe its inner workings to make debugging easier, while an auditor looking into the fairness of an algorithm may be satisfied with explanations that focus on the output.

Right now, XAI is more of a shared goal of creating transparency in the AI system than a well-defined practice.\textsuperscript{19,20} As we explain in the next section, meeting the goal of explainability takes working across the entire development cycle of an AI system as well as in its deployment.


\textsuperscript{19} Reza Shokri, Martin Strobel, and Yair Zick, Privacy Risks of Explaining Machine Learning Models, arXiv:1907.00164v4, 2019

\textsuperscript{20} Ibid. Gilpin
Both the AI community and industry are coming to see the need for more transparent AI systems. As AI solutions are evolving past contained proofs-of-concept to deployment at scale, they recognize the importance of prioritizing XAI to satisfy adoption, to power effective human-AI collaboration and to satisfy audit and regulatory needs. An important challenge for applied scientists in the field is how to meet all of these needs.

As XAI is a dispersed field, there is no clear definition of quality for an explanation. Some even argue against investing in post-hoc explanations of black box models for decisions like credit scoring. However, it is clear that any evaluation must be based on the desired output and therefore any method needs to be applied in a user-centered approach. Furthermore, relying on purely technical solutions to provide explanations ignores the complexity of implementing XAI in complete systems that integrate within business processes.

Most current AI practice focuses on performance, with explainability dealt with as an afterthought. Designing for explainability requires evaluating the needs for transparency in an AI system and taking them into account from the initial steps of building a solution to the system rollout.

Both the AI community and industry are coming to see the need for more transparent AI systems. As AI solutions are evolving past contained proofs-of-concept to deployment at scale, they recognize the importance of prioritizing XAI to satisfy adoption, to power effective human-AI collaboration and to satisfy audit and regulatory needs. An important challenge for applied scientists in the field is how to meet all of these needs.

Instead of a focus on performance over explainability, AI scientists should work towards both.

A Multi-Faceted Solution
The Element AI XAI team has worked with several external clients to integrate actionable insights provided by XAI capabilities to support their processes. We have recently worked with a tier-1 manufacturer to build a multi-XAI component workflow on top of their existing visual anomaly detection model. The manufacturer needed to explain accuracy, the confidence in predictions as well as the in- and out-of distribution operating ranges for plant operators and customers to trust the results. Using multiple techniques together, we worked with them to develop a robust capability for inspecting the decision-making of their models that enabled them to reduce their false positive errors by more than 90%, while maintaining nearly 0 false negatives.
As XAI is a dispersed field, there is no clear definition of quality for an explanation. Some even argue against investing in post-hoc explanations of black box models for decisions like credit scoring.\(^{21}\) However, it is clear that any evaluation must be based on the desired output and therefore any method needs to be applied in a user-centered approach. Furthermore, relying on purely technical solutions to provide explanations ignores the complexity of implementing XAI in complete systems that integrate within business processes.

Designing for explainability needs to make use of an ensemble of practices. It means constructing the AI system with an evaluation of the explainability needs in mind, including how the data is collected and processed, which AI model is chosen, how it is trained, how it picks a decision to recommend, how the interaction with the end-user is designed and how it is deployed. Successfully implementing the appropriate explainability characteristics of a system across its entire cycle of development requires expertise that goes beyond the highly specialized expertise of an AI scientist and extends to the spheres of governance, policy, product design and system engineering.

Explainability is an essential characteristic of good machine learning solutions, and can be addressed in many ways. Current approaches have provided value on many points, but as we move to a world of integrated human-AI collaboration, research will need to embrace multidisciplinary approaches that integrate the needs and ideas of multiple fields.

---

Element AI is one of the few companies to have a full-time dedicated XAI team. We have two driving goals: first, to help our clients solve problems of transparency in current and upcoming AI systems by developing and integrating the existing capabilities from the field of XAI. Second, to push the boundaries of research to develop novel XAI approaches that can deliver more actionable outputs of AI solutions.

Element AI’s XAI team is composed of a multidisciplinary group of scientists, designers and policy experts. Designing explainable AI systems isn’t easy. It can require acting anywhere in the AI system’s value chain from from modelling and experimentation to deployment and monitoring, as well as the system engineering necessary to evaluate and correct a model once it is in production.

There isn’t one switch to flip to include explainability in an AI system, and integrating it can be a costly and resource-intensive exercise. However, taking a holistic approach to the problem, we are working to develop the tools, interfaces and approaches to make actionable, transparent AI accessible for a variety of users from ML developers to non-technical users interacting with the AI system.

Quickly solving transparency problems in existing AI solutions

Our team is constantly evaluating the large and growing body of XAI literature, which appears across different domains and titles such as “interpretability”, “transparency”, “trustworthiness” and “intelligibility”. As most XAI methodologies are poorly tested for their impact on helping users with specific tasks, we invest heavily on validating methods with user studies and tailored interfaces before proposing them in solutions. This allows us to move quickly and confidently with our chosen approaches for transparency problems in AI systems currently, or soon to be, in production.

A salient example of how AI isn’t just a “flip of a switch”. When working with feature attribution on images for a retail manufacturer, we realized that out-of-the-box methods from academia often gave different explanations for the same model; and that the heat maps used to serve the explanation could be impractical for non-technical users. By overlapping multiple techniques and using contextualized visualizations that guide the user to the points of agreement and disagreement, we have aimed to bypass these technical limitations and provide actionable information to the end-user.

---

XAI literature appears across different domains and titles such as “interpretability”, “transparency”, “trustworthiness” and “intelligibility”

Building the explainable AI capabilities of tomorrow

Over the longer term, our scientists are looking towards the future and building out new methods for explainability that support trustworthy and practical AI solutions. Bold and impactful research is at the core of Element AI’s mission, and we recognize designing for explainability as a complex undertaking that requires working from multiple expertises. Our world-class research pipeline helps us develop and identify state-of-the-art innovations, turn them into testable prototypes, and use the best ideas to power products that deliver real business value. XAI is no exception, and we have big plans for the future.

AI has incredible capabilities and is already driving business value in industries like financial services. XAI can help drive user adoption and business value in a number of ways, helping address the core problem of transparency and trust as well as regulatory compliance, bias and time to market for AI solutions. As the capabilities and applications of artificial intelligence spread, XAI will become a key driver of business value for the companies that seize the opportunity.
ACKNOWLEDGEMENTS

This discussion piece originates from an internal report written by applied research scientist Bahador Khaleghi. It was updated and adapted for the financial services sector with contributions from Alexei Markovitz, Ben Crestel, Massimo Caccia, and Richard Zuroff, and edited by Simon Hudson.