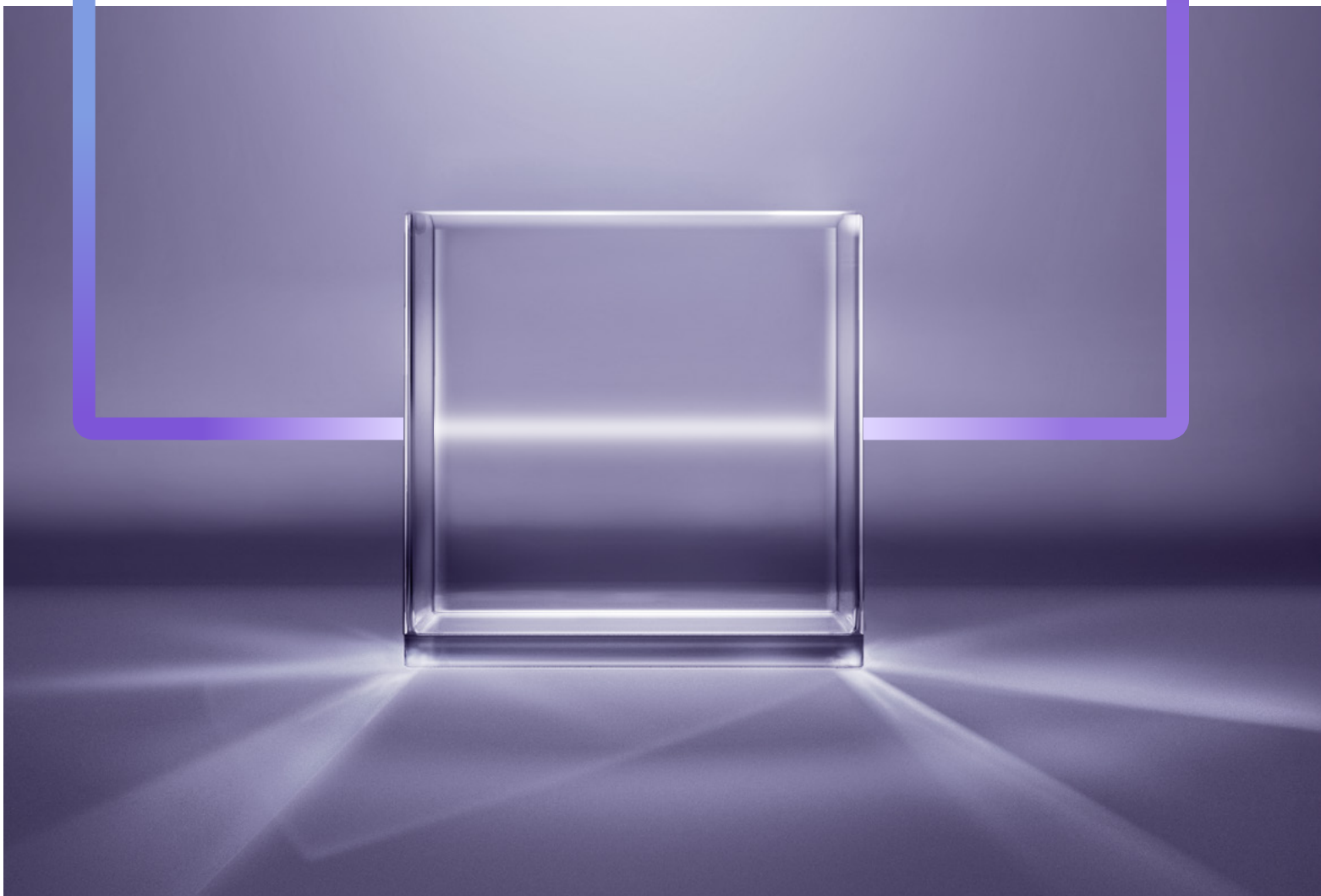
ELEMENT AI

For more information
sales@elementai.com

Opening the Black Box

Driving business value with AI



Opening the Black Box

Driving business value with AI

Artificial intelligence is changing the world around us. Every day, AI systems are buying and selling millions of financial instruments, searching cases and building precedent for lawyers, and helping big companies identify and fight off sophisticated cyberattacks. As adoption spreads, it is not enough for AI systems to perform well. We need to understand how they work.

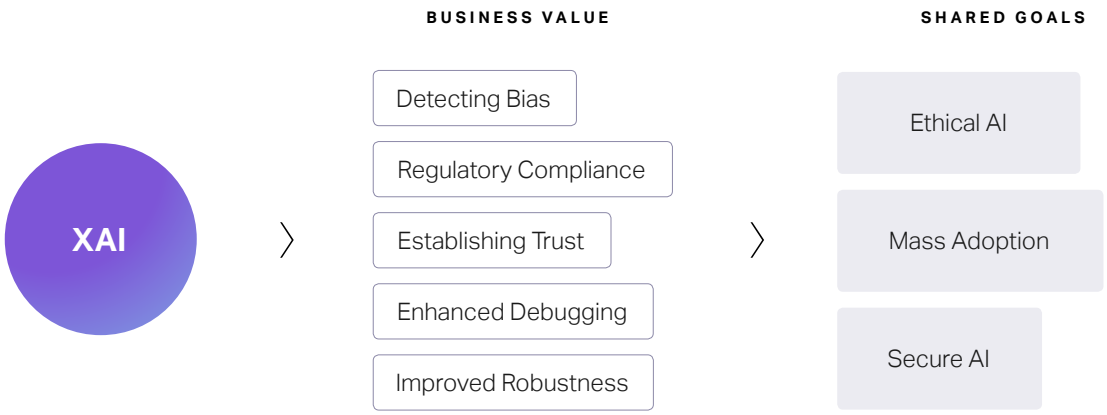
The challenge for AI, unlike previous technologies, is that how and why it works isn't always obvious. Many of the advanced machine learning algorithms that power AI systems are inspired by the human brain, yet they lack the human ability to explain their actions. We can see how an AI system acted in response to a certain input, but we don't always know why.

Thankfully, there's an entire research field working towards describing the rationale behind AI decision-making: Explainable AI. XAI, as it is known, is experiencing a new wave of interest as modern AI systems have demonstrated performance and capabilities far beyond previous technologies. And XAI could be a key driver of business value for the companies that are putting AI to work.

The business value of Explainable AI

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

Explainability effects



Technology has always driven change in the business world. AI is only the latest in a series of profound shifts in the way we do business, the most recent of which have been brought on by the development of modern computers. In every case, one thing is true: in order for technology to be useful, it has to be put to use. Developing a solution is one thing, but the biggest challenge is getting your user community to adopt it.

To drive AI adoption, explainability is key.[1] There is always a transition period in which the user community validates the capabilities and limitations of any technology. On a personal level, we want to work alongside systems we can trust. Explainable AI will go a long way towards building credibility for the AI technology if we know both why and how it is reaching its decisions.

It's not only in user adoption where XAI can drive business value. Other motivations for XAI include regulatory compliance, detecting bias, and

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

[1] Khaleghi, B., 2018. Breaking down AI's trustability challenges. Element AI. [online] Available at: https://www.elementai.com/news/2018/breaking-down-ais-trustability-challenges [Accessed June 18, 2019].

The future of compliance

Data-rich, regulated areas such as healthcare and financial services offer some of the most promising near-term applications for AI decision-making. Right now, the law requires explainability in only a small subset of decisions, even within those regulated industries. There is immense value to be unlocked by integrating XAI capabilities into an AI system to automate that subset of decisions. And beyond that small number of regulated decisions, there is much opportunity for AI optimization in less sensitive areas.

Governments and regulators are beginning to study AI and its impacts, and explainability could play a part in future compliance with emerging regulation. Lawmakers in the United States have introduced the Algorithmic Accountability Act, which calls on large companies to check their algorithms for bias, and EU politicians, who have led the way on privacy legislation with the GDPR, have discussed similar accountability measures in the trade bloc. Pursuing XAI now could open up the opportunity for a wider number of AI applications in regulated industries and beyond in the future.

XAI could become a competitive differentiator for those who integrate it into their AI systems.

protections against 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 their fair treatment of protected classes such as race and gender. 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 detect pneumonia in X-ray images, where subtleties in the data such as image compression can change the diagnosis, and those used to give out bank loans, which could end up incorrectly denying loans to those from minority communities.[1,2] In each case, the AI systems were making decisions based on features within the data that would or should be irrelevant to a person facing a similar choice.

Lastly, XAI can also deliver business value through speed. An AI system built with explainability in mind would have a shorter development time (and time-to-market) by making it easier to debug, enhance and tune the AI models that power it.

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.

[1] Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J. and Oermann, E.K., 2018. Confounding variables can degrade generalization performance of radiological deep learning models. arXiv preprint arXiv:1807.00431.
[2] Waddell, K., 2016. How Algorithms Can Bring Down Minorities' Credit Scores. The Atlantic. [online] Available at: https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/ [Accessed June 18, 2019].

Explaining Explainable AI

Modern AI has capabilities that match and exceed humans in certain tasks. We don't always know how it works. Explainable AI is an evolving research area that aims at making machine decision-making understandable to humans.

XAI is the effort to provide the reasoning behind an AI model, both its output and its inner workings. 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 clear what happened in the middle.

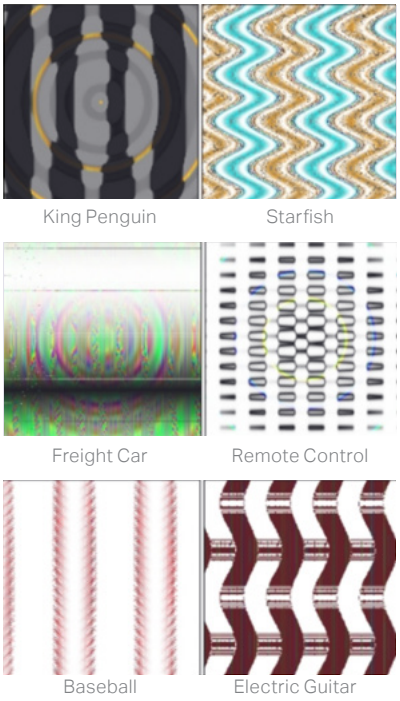
Explainable AI is about figuring out those black boxes and that missing middle: how data is being processed by the model, and how that affects the model's output.

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.

One easy example to aid in understanding the definition of XAI is in image recognition, where modern AI made its first big gains back in 2011 and 2012.[4] To the right are several images unrecognizable to a human that were nevertheless misclassified by a state-of-the-art AI algorithm.[5]

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 labeled a remote control. Though the AI is able to recognize these 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.

Right now, XAI is more of a goal than a practice. There are a few established methods and a wide variety of approaches, only a few of which have been proven useful. We don't even have an agreed-upon definition, in part because most explanations now are generated based on questions after the fact. These "why?"



These images are unrecognizable to humans, but state-of-the-art deep networks identify them as familiar objects with over 99% certainty.

[4] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
[5] Nguyen, A., Yosinski, J. and Clune, J., 2015. 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).

a) Husky classified as wolf

b) Explanation

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

While AI models have become much better at certain tasks, they still do not reason like we do.

questions vary based on the asker: a developer creating an AI model might prefer explanations that describe its inner workings to make debugging easier, while an auditor looking into the fairness of an algorithm may prefer explanations that focus on its output.

With some image classification models, it's possible to demonstrate which pixels or sections of an image are most important to the given output, a technique known as feature attribution. Feature attribution is currently one of the common approaches to XAI, and this kind of explanation can help detect spurious correlations: in one oft-referenced example, an algorithm mislabeled an image of a Siberian Husky as a wolf.[6]

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.

[6] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. Why should I trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). ACM.

Explainability by Design

Most current AI practice focuses on performance, with explainability dealt with as an afterthought or ignored completely. Explainability by Design, where the entire AI development pipeline is designed with explainability in mind, should be the way of the future.

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

Within the AI community, there is still not enough recognition of a simple fact: people need to trust AI in order to put it to use.[7] Most science is focused on benchmarks and improving performance on key datasets. Instead of a focus on performance over explainability, AI scientists should work towards both.

There is currently no shared set of standards or benchmarks for measuring the quality of explanations. Researchers often measure quality indirectly through proxies such as the explanation's agreement with human intuition, or its satisfaction of some mathematical property — consistency, robustness — desirable for explanations.

The XAI literature is mainly focused on post-hoc explainability methods such as feature attribution, where the goal is to extract explanations from an existing model after the fact. It's challenging to judge the quality of these explanations, because they are often evaluated based on their intuitive appeal. We might like and understand an explanation, yet it may not reflect the actual behaviour of the underlying model.

Explainability by Design is the way of the future. It means constructing the entire development pipeline with explainability in mind, including how the data is collected and processed, which AI model is chosen, how it is trained, and how it is deployed.

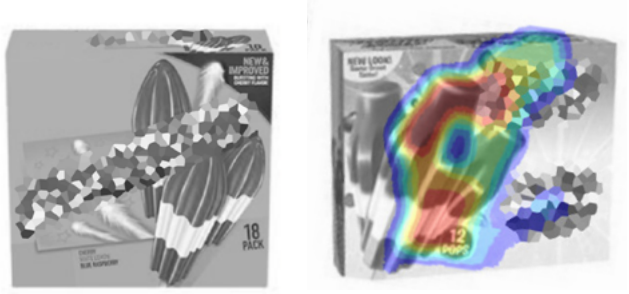
Current choices for XAI by Design models are limited and can offer lower predictive performance. A promising new direction is explainable models that come with optimal performance guarantees, though applicability of such methods is limited at this point. There's a significant need for more and better XAI research, providing more capable methodologies with the best of both worlds.

[7] Khaleghi, B., 2018. A missing ingredient for mass adoption of AI: trust. Element AI. [online] Available at: https://www.elementai.com/news/2018/a-missing-ingredient-for-mass-adoption-of-ai-trust [Accessed June 18, 2019].

Explainable AI at Element AI

At Element AI, we are one of the few to have a full-time dedicated XAI team. We have two major goals: first, understand the existing capabilities of current XAI and integrate the state of the art into our product portfolio, and second, pushing the research forward and developing novel XAI approaches.

Our explainability team has already integrated some explainability features into an existing solution as a proof of concept for the value of XAI. We developed an efficient solution for detecting copycats of retail products based on the appearance of the packaging. The problem was that our client could not rely solely on the predictions, the outcomes of our AI system, to take legal action. To solve this, we changed the system so that it could provide reasoning for its prediction, highlighting the parts of the packaging that the model deemed to be most similar to the original product.



Original Brand Copycat Brand

Attention map shows the regions that had a high contribution in the model's decision.

XAI isn't easy. As previously stated, it's a goal and not a practice. 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 in some cases and impossible in others. Yet we are working in the short and long term with Explainability by Design as our goal, so that eventually it becomes second nature to researchers within Element AI and beyond.

In the short term, our scientists are ingesting the large and diverse body of XAI literature, as it is being studied in different domains and under different

Our dedicated XAI team is working to push science forward.

While AI models have become much better at certain tasks, they still do not reason like we do.

names such as interpretability, transparency, and intelligibility. The number of publications is rising as the field expands, and there is much to learn. We have realized that common XAI methodologies, such as feature attribution, are not adequate when it comes to explaining models with complex output, such as one trained to identify letters or a sequence of words given a partially cropped input image.

Over the longer term, our scientists are looking towards the future and building out new methods for explainability. Bold and impactful research is at the core of Element AI's mission of helping people and organizations work smarter. 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.

We're also working on XAI because it helps build trust in AI, which drives adoption, which drives business value. We're working on it for the good of AI itself, which needs explanations to help move the science forward. And for AI to be put to use in the widest possible way, we need to make it accessible and compliant in the widest number of applications.

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

Montreal
6650 Saint-Urbain
Suite #500
Montreal, QC, H2S 3G9
Canada
+1 (514) 379-3568

Toronto
296 Richmond St. W
Suite #100
Toronto, ON, M5V 1X2
Canada

London
2 Eastbourne Terrace
London
W2 6LG, United Kingdom

Seoul
Dreamplus Gangnam
311 Gangnam-daero
Seocho-gu, Seoul
Republic of Korea

Singapore
60 Anson Road
Level 17
079914, Singapore

CONTACT US
sales@elementai.com
+1 (877) 670-8843 #102

ACKNOWLEDGMENTS
This report was written by Applied Research Scientist Bahador Khaleghi, edited by Peter Henderson, with additional contributions from Collin Mechler and Simon Hudson.