Can you trust AI?

Chris Brannigan and Joe Faith explore the necessities of explainable AI

Across autonomous vehicles, healthcare and criminal justice, artificial intelligence (AI) is supporting high-risk, high-consequence human decisions. In healthcare, AI systems perform as well as human physicians on cognitive tasks in radiology, pathology, dermatology, and ophthalmology. The 'rise of the machines' appears inevitable. Breathless headlines estimate that AI could save the global banking industry over $1tn[1] by 2030.

However, progress for AI in high-risk decision environments such as financial crime presents significant new challenges. Who is accountable for AI when it goes wrong? How can a complex AI system be transparent in its working and operation? How did the machine make the decision?

It comes down to an issue which is at the very core of banking and the financial system – trust. How do you trust the machine?

In this article we will explore this question of *explainability* – a hot topic in AI. We will see how generating human-interpretable AI outcomes is vital in order to establish trust and for improving the performance outcomes of AI. There are answers, some of which may be surprising. Humans have a big role to play in the tale of the machines.

Machine – explain thyself

The AI that we refer to here is not a generalised self-learning intelligence. That remains the stuff of science fiction. We refer to the broad sub-set of machine learning algorithms that, when given a set of inputs about a domain problem, can make predictions after learning on many historic examples in that domain.

The most prominent examples are 'deep learning' algorithms. While these have been the foundation for many of the high-profile breakthroughs in AI performance, the downside is that it has become increasingly hard, and sometimes impossible, for humans to understand how individual decisions are reached inside these black boxes.

Financial service firms have historically developed and deployed systems that are primarily rules-driven. These are often subject to laws, regulations, regulatory guidance, policies and widely-accepted practices that have set stringent standards in operating systems. These systems and models are often rigorously reviewed and audited by multiple parties, including regulators.

For Nishanth Nottath, Global Head of Transaction Monitoring at HSBC, the question that lies at the core of any review is "did the system do what it was expected to do?" (see also p.12-15). State of the art AI operates on non-linear relationships, uncertainty and probability. This is the first problem.

For risk executives and regulators raised under the dark clouds of the global financial crisis and ongoing ►

Figure 1: Example of AI explained outcome utilising rules and anchors method

```
IF Country = United-States AND Capital Loss = Low
AND Pension = True AND Relationship = Family
AND Married AND 28 < Age < 37
AND Location = Urban AND High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary > $50k
```

Figure adapted from Ribeiro et al (2018)

financial crime penalties, this urgently leads to the questions of 'how' and 'why' decisions were made.

Ensuring trust in this context means developing AI financial crime solutions that can explain how they reached a decision to analysts, executives, auditors, regulators and customers.

When AI goes bad

Like humans, AI predictions weigh up uncertainty and probability of outcomes. Like humans, even with lots of experience (and good data), mistakes will be made.

As Dev Odedra, a financial crime consultant who sees very obvious parallels in the approach taken for both humans and machines, explains: "When training AML investigators, one of the first things I used to teach them was 'rationale is king': show your working out. It is better to make the wrong decision with the right rationale than to make the right decision with an unexplainable rationale".

In the high-consequence world of financial crime, a black-box AI provides the bank with a poor defence when mistakes are inevitably made. For regulation and audit, explainability is a requirement for AI. This capability provides an enhanced method for AI model performance testing and diagnostics.

For example, an AI system may determine that an alerted set of transactions were not suspicious. An explainable AI may have determined that the counterparty in the transactions was legitimate, had a valid business relationship with the focal entity and the amount was in line with business expectations. The AI may provide evidence to support these judgements: legitimate company website; business registered as expected; directors checked; known ownership structure; long standing business relationship; nature of business expected; no adverse news; transactions in line with similar businesses in sector.

The consistency and clarity of the explanation is very useful for an audit that may find that the determination was incorrect. The explanation provides a clear diagnostic pathway:

- Which judgement(s) in the justification was/were incorrect?
- Was the algorithm at fault?
- Was data unavailable or not adequately labelled?
- Was this a new type of suspicious activity not previously captured by the bank?

Explanation is a powerful tool for AI solution developers to work with banks to continually diagnose and enhance performance. This is needed, as financial criminals are constantly evolving.

Malicious data

As banks develop governance and processes to ensure the highest data quality for AI algorithms, they may be faced by a familiar enemy: fraud. Malicious actors may insert information to bias and fool AI systems for the purposes of financial crime.

This emerging threat was convincingly demonstrated by Harvard scientists, Finlayson and Beam, who were able to methodically mislead AI computer vision systems to make false clinical diagnoses in three different medical domains: assessing retina scans; chest x-rays; and scans of moles for signs of skin cancer. This was achieved by changing a single pixel in a scan, undetectable to human experts.

Progress for AI in high-risk decision environments such as financial crime presents significant new challenges

As AI systems become more sophisticated, they will gather evidence and make judgements on many sources of internal and external financial crime data. Financial criminals are sophisticated and will seek to take advantage of this.

In these instances, AI explainability is a vital tool for regulatory defence, model error identification, diagnostics and enhancement. So, how do banks make explainable AI a reality?

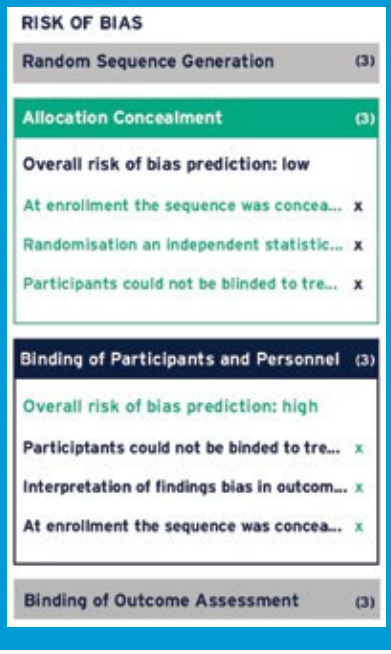
Explainable algorithms

The AI development community is feverishly experimenting with new algorithms to produce human-interpretable AI outputs. A few high-profile methods include:

- **Simple Algorithms** – Algorithms such as regression, decision trees, and graph methods are easier to explain than deep learning (see below), although there is an impact on accuracy. This approach works well in narrow, well-defined problems of cognition.
- **LIME** (Local Interpretable Model-Agnostic Explanations) is a very popular approach and is ►

Box 1: Robot-Reviewer

Robot-Reviewer enables researchers to upload PDFs of scientific research papers.
The system can then read and analyse these documents and produce a detailed assessment of biases that exist within the experimental design.
The solution applies the industry standard 'Cochrane Risk of Bias' methodology as a framework and has built models to assess and explain within this framework.



easy to implement. It utilises linear algorithms to 'probe' the black box. However, if the problem is better solved using a non-linear method, then there will be a trade off in model accuracy.

- **Rule-based anchors** – These present the decision explanation in the user-friendly form of an 'If-Then' statement. This method is one to watch for narrow compliance tasks (see **Figure 1** for an example).
- **Layer-wise Relevance Propagation (LRP)** – What the name lacks in elegance, the technique makes up for in power and potential. In short, LRP takes your favourite neural network and does it backwards.[2]
- **Bayesian Deep Learning (BDL)** has the advantage of applying a powerful algorithm to maximise model performance. It is possible to map what feature led to what decisions and the relative importance of it. This approach shows lots of potential. At Caspian, we are big fans of LRP in combination with Bayesian approaches, which we currently utilise in our Financial Investigator Platform (FIP).

However, algorithms are not the whole answer to explainability.

How do humans explain?

Other approaches start by defining the thinking processes utilised by human experts when undertaking specific complex cognitive tasks, such as medical diagnoses. They then use these cognitive maps as a design architecture. They develop models to replicate different components within the cognitive map and utilise the architecture to make decisions and to explain the outputs.

These have been applied to very promising effect to automate and augment complex human activities (see **Boxes 1** and **2**).

Humans make a comeback

It's not all over for homo sapiens. It never was. The extreme view of full AI autonomy is now being rapidly revised to put humans in the loop.

AI must be explainable if it is to be trusted for high-consequence decisions. Humans are vital in providing a cognitive framework for AI explanations. These interpretable predictions can then be assessed by human analysts to precisely diagnose errors and feedback into the AI.

The activity of training machines should be a collaborative endeavour between humans and machines. Explainable AI should automate and augment high-consequence tasks to support well-trained human expertise.

The AI explanation barrier is solvable and great progress is being made. If successful, the enormous potential of AI in financial services may be fulfilled.

Chris Brannigan is CEO and Founder of Caspian and Joe Faith is Caspian's Chief Product Officer

1. https://next.autonomous.com/augmented-finance-machine-intelligence
2. Credit: Brendan Whitaker

Box 2: FIP (Financial Investigator Platform)

FIP is a solution developed by Caspian that automatically investigates AML Transaction Monitoring Alerts and conducts KYC Account Reviews in financial crime. The solution outputs human-readable explanations that detail judgements made and supporting evidence. The output report provides direction for human analysts to check and complete any remaining investigation.

The solution has captured a cognitive map of how expert investigators think and act when analysing different financial crime problem domains, for example, the key questions that an investigator will seek to answer when establishing that an activity is in-line with a counterparty's nature of business. Multiple AI models then make predictions and formulate explanations within that expert architecture.