Please explain yourself to us

__Gren Manuel__ discusses the tricky issue of AI explainabliity and outlines why regulators want to ensure that banks' AI systems can demonstrate how they got their answers

If artificial intelligence is to fulfil its promise in financial services, it must solve a fundamental problem: explainability. This concept may seem like an arcane technical issue, which the AI team should solve as a technical exercise. In reality, it is a deep-rooted challenge for senior managers and regulators.

The issue has been brought into sharper focus by the excitement around the latest wave of AIs known as Generative AI, typified by ChatGPT, which can do human-like tasks, such as producing text and writing computer code. All these are driven by 'foundation models' – general-purpose AI models based on neural network technologies that, in some ways, mimic the human brain. Simply put, they operate like a sophisticated and large-scale 'auto complete', which responds to user prompts – prompts that can now include uploading a picture of the contents of your fridge and asking for recipe ideas.

Janet Adams, Chief Operating Officer of SingularityNet, a decentralised AI marketplace, who held senior positions in conduct and risk at three major UK banks, says of these technologies: "They're brilliant, but they're completely opaque and completely unexplainable. And that's what really holds them back from being usable in financial services."

Foundation models 'learn' from ingesting gigabytes of data and their behaviour emerges from complexity rather than being explicitly coded. Although their neural network technology takes inspiration from the brain, their output is more like an inspired human hunch than high-level cognition.

In some applications, this may not matter. The UK government's AI regulatory framework notes that some important decisions made by humans also defy explanation. And there are many

examples where outcomes matter more than explanations: for instance, some pharmaceuticals are routinely prescribed for severe medical conditions where researchers still do not understand why they work.

But financial services is different. The benefits of the smallpox vaccine in the eighteenth century far outweighed any risks, even if no one could explain how it worked. For financial services, knowing what you're getting yourself into is an essential part of making it work. It helps ensure both consumer protection and financial stability, and it's why some products are only available to 'sophisticated' investors.

> *People fail to appreciate that explainability is important because it is good business*

In a Bank of England survey in 2022, firms put explainability and interpretability as their top-ranked risk because of potentially bad consumer outcomes and subsequent reputational and legal risk. But there is also a potent risk to firms themselves, and to the wider economy, of blackbox financial products, as the infamous 'CDO cubed' of the financial crisis of 2007-08 made clear.

The emphasis on explainability will be locked into place as the UK installs its first AI regulatory framework – and the focus is on consumer protection. The 2023 White Paper proposes a non-statutory framework of five principles, one of which is "appropriate transparency and explainability". To be given teeth, these will be handed to regulators across the UK economy, including the Financial Conduct Authority (FCA) and the Prudential Regulation Authority, to be turned into industry-specific regulation.

Operating at multiple levels

But to make things more complex, explainability has to operate at multiple levels. Local-level explainability is the industry buzzword for explaining how an individual decision was made. Most lawyers reckon that under the Data Protection Act 2018 (which, in effect, keeps the UK compliant with the EU's General Data Protection Regulation post-Brexit), anyone subject to computer decision-making has a legal right to an explanation. This means any company's AI model that is accepting or rejecting mortgage applications, for instance, has to be able to provide the applicant with an explanation of its decision.

Global-level explainability is the phrase used for explaining how an AI system works in general, such as which factors are most important in decisions and how they are processed and weighted. Only by looking at this would a firm be able to reassure itself that it is meeting the FCA's Consumer Duty requirements, such as avoiding harm, bias, or selling products that customers cannot afford or offer them no value.

These are the regulatory reasons for requiring explainability, but there are managerial reasons, too. Clara Durodié, Chief Executive of Cognitive Finance Group, a research and advisory firm focused on AI in financial services, says: "It's not just regulations. People fail to appreciate that explainability is actually important because it's good business."

Hani Hagras, Chief Science Officer at banking software provider Temenos and a professor at the Artificial Intelligence Research Group at the University of Essex, provides a compelling illustration. An AI model could be fed historical data about customers closing their accounts and then predict with high accuracy who will likely leave in the next three months. But, without explainability, it can't explain why a customer is leaving, meaning the bank can't easily work out how to retain them. Hagras says: "It is explainability that will answer the question of why the customer was not happy with their bank."

More broadly, he says, explainability is critical to ensuring that customer service driven by AI is effective. "You need to make sure that you provide the same kind of output you're going to get from a [human] relationship manager, and only with explainability can you provide this. Without this, AI will not realise its full potential."

Explainability also helps identify sources of bias (see 'The Devil is in the Data' for more on this). Without reassurance that the AI is operating without bias, financial firms risk having to route any sensitive application through a separate, non-AI process, which will increase costs and reduce the returns on the AI investment.

> *Without explainability, your AI project will be a big money hole, and you risk censure and failure*

The best solution would be technical. One much-researched option is to retrofit explainability. Many examples are fed through an AI process and other algorithms (SHAP and LIME are well-known examples) back-calculate rules and factors that would generate similar results. It sounds like a solution, but any output that is understandable to a human being is unlikely to replicate the exact output of the AI fully. And making it more accurate may make the explanation increasingly complex and impenetrable.

An AI arms race

The way forward would be for new AI technologies to come onto the market that are as powerful as neural networks but with inbuilt explainability. Temenos has patented AI technologies that produce models that can be understood, explained, analysed and augmented by business and lay users. These can be applied to banking tasks such as payments, wealth management and preventing and detecting financial crime and money-laundering.

But competing technologies will need help keeping up with the power of neural networks. Google, Meta Platforms (owner of Facebook) and AI specialists such as OpenAI (backed by Microsoft) are in an AI arms race, pouring billions of dollars into developing and training new models based on neural networks. With this investment, and their ownership of the hyper-scale cloud platforms needed to carry out the research, they will likely improve faster than other technologies.

Durodié says managers should ask two questions: 'Are any of our AI tools doing anything we cannot explain?' and 'If any of our AI tools aren't explainable, what risks are we facing?'

For some tools, the answer to question one may be 'yes', but the application may present no problems. An AI chatbot that answers questions about the interest rates on savings products doesn't need explainability. Neither does an AI tool deployed in the network to improve security.

When it comes to question two, a balanced assessment is required. Issues such as bias and accessibility must be considered, and reputational and regulatory risks must be assessed.

While at a global bank, Adams ran a 10-week Introduction to AI course and she says courses such as that are essential if banks are to get to grips with AI – both the potential and the problems. She believes a full understanding of explainability is vital for any AI project. "Without this," she says, "your AI project will be a big money hole. And you risk censure and failure in all kinds of ways."

Overall, the issue of explainability adds significant complexity to using AI in financial services. Anyone designated as a material risk-taker under the FCA's Senior Managers and Certification Regime must ask tough questions when AI solutions are proposed. These are different from the questions that had to be asked in the past when third-party technology was deployed inside the institution. And, while it's true that UK regulators are looking to oversee services provided by 'critical third parties', such as technology companies, the buck will stop with the financial services firm. ■

Gren Manuel *has been European editor for Dow Jones Newswires, European executive editor of The Wall Street Journal, and editor of Financial News. He now works as an editorial and media consultant*