UK FINANCE

EY
Building a better working world

TRUST, CONTEXT AND REGULATION:

Achieving more explainable AI in financial services

September 2020

TABLE OF CONTENTS

1. FOREWORD

The use of artificial intelligence (AI) in financial services continues to grow. As the Financial Conduct Authority (FCA) and Bank of England identified in their 2019 research note on machine learning, industry take up is increasing and use cases are expanding out from the back office and into customer-facing applications.[1] We also see a shift towards more complex models.

This expansion brings many opportunities for the industry to improve efficiency, better manage risk and provide exciting new products and services to customers. However, to take full advantage of this opportunity – there needs to be trust. As with all innovations, ethical considerations must keep pace with technological development.

Building trust requires transparency and communication. Indeed, this is a topic of growing regulatory and government interest in many countries. Transparency and communication with customers have long been key considerations for financial services but AI will require new approaches and techniques if explanations are to be meaningful. Effective explanations will also require a degree of subtlety; given the huge potential range of use cases, close attention to the context of each will be key.

Alongside this, consumer education as to how and why AI is being used is increasingly important.

Achieving effective explanations will require firms to have a clear AI strategy and robust governance, and to engage effectively with colleagues from a range of functions, including data science, compliance, audit, business and senior management, and even ethicists. It will also require ongoing work, with limits yet to be resolved in the state of the art of explaining AI and with 'best practice' sure to evolve. More research and thinking will be needed, not just from firms but also from regulators, government and think tanks.

We hope that this paper is a helpful contribution to this developing area of technology and policy.

Sajedah Karim
Partner, Financial Services RegTech and Compliance Transformation Leader, Ernst and Young LLP

Hannah Gurga
Managing Director for Digital, Technology and Cyber, and Chief of Staff, UK Finance

1. https://www.fca.org.uk/publications/research/research-note-machine-learning-uk-financial-services

2. INTRODUCTION

a. About this paper

The challenge is not only providing better explainability, but also knowing why and when an explanation is needed, who is accountable for providing it and how to ensure information needed for the explanation is available and accessible.

While AI adoption in the financial sector is growing rapidly, there are concerns that AI could exacerbate certain risks and that AI models may not always fit cleanly within existing regulatory frameworks. The increasing use of AI-based systems for real-world applications raises ethical and regulatory questions about the harm that could be caused through unintentional failures and the need to check that existing regulatory regimes effectively cover developing practices. As such, AI-based systems are increasingly attracting the attention of regulatory agencies and wider society, including media and civil-society groups.

Models used in AI-based systems are becoming more sophisticated and effective, yet important issues remain to be resolved. Concerns relate in particular to accountability, safety, unfair algorithms, bias in data, and the application of models that are not appropriate to the business context – as well as ensuring that the use of models provides value to consumers.

One important consideration for firms expanding their AI capabilities is that AI-based systems can, depending on the model, be 'black boxes'. Such systems perform functions and make decisions that cannot be easily accessed or interpreted by human beings and can be costly to monitor. Without the ability to clearly interpret, explain and monitor AI outputs and models, it can be difficult to ascertain that a decision has been made in line with a firm's intention and to communicate effectively with impacted individuals.

This may be acceptable for some use cases, but businesses operating in the heavily regulated financial services sector need to be sure that they can achieve a level of transparency and explainability that is aligned to regulatory expectations, and which increases customer trust, while also striking a balance with system accuracy. This is important for how firms communicate decisions to their internal stakeholders and business, as well as to end customers, and has implications for how they consider and manage risk and maintain the resilience of their services.

This paper will focus on why and to what extent explainability of AI outputs is needed, the challenges to achieving this and potential ways to apply the latest guidance. It will also provide technical financial services use cases to explore potential approaches to different types of explanations, according to the context and type of model. The paper considers not just common existing uses of AI, but also emerging or possible uses.

BOX 1 – DEFINITION OF AI

Artificial intelligence (AI) is an umbrella term for a range of algorithm-based technologies that are designed to mimic human thought to solve complex tasks. In some instances, AI is thought to include traditional form of statistical analysis, such as linear regression.

In other instances, the term is reserved for newer, self-adapting techniques such as machine learning and deep learning. This paper takes a broad view, in line with the approach of the Information Commissioner's Office (ICO), though some considerations will be more relevant for more sophisticated, adaptive technologies.

We note that context is important; regulations and rules may require tight definitions, focused on the specific technologies of interest.

Decisions made using AI are either fully automated or include a 'human in the loop' that is involved in making each decision, though the exact role can vary.

Within financial services, emerging use cases include identifying suspicious transactions such as money laundering, and evaluation of a customer's creditworthiness using diverse datasets.

b. The regulatory environment

Financial services firms have long needed to explain their decisions to customers, regulators, investors and stakeholders. Regulations such as the European Union's Revised Payment Services Directive, the Markets in Financial Instruments Directive and, in the United States, the Equal Credit Opportunity Act and the Fair Credit Reporting Act, all require firms to explain their services.

In the UK, firms must consider rules imposed by the FCA such as MCOB 4.7A.23A which requires firms to explain why a more expensive mortgage product has been chosen when a cheaper option is available. They must also consider the extensive disclosure requirements within Conduct of Business Sourcebooks, which set out the information that must be provided to customers, how it should be presented and when it should be provided.

There are also industry standards that are relevant to certain business lines, such as the Standards of Lending Practice.[2] Often, where conduct risk-related reviews are performed, extensive discussion is required to determine how an outcome was reached and whether it was fair and unbiased.

More recently, in response to the increased interest in AI-based systems, many public and private organisations have established AI principles, covering guidance on safe AI adoption and raising challenges around potential AI regulation. In general, these highlight the importance of considerations such as fairness, ethics, transparency, explainability, accountability, resilience and trust.

Public sector publications indicating the likely future direction of regulation on the use of AI include the European Commission's Principles for Trustworthy AI and the 19 February 2020 European Commission white paper 'On Artificial Intelligence - A European approach to excellence and trust' [COM/2020/65 final] on a potential AI regulatory framework.

2. https://www.lendingstandardsboard.org.uk/the-slp/

A prominent example of a set of AI principles is the OECD AI Principles, which were adopted by the G20 nations (see Box 2). One principle, common across many sets of AI principles, is that AI systems need to be in some way transparent or explainable, or both.

BOX 2 – OECD AI PRINCIPLES

The OECD identified five complementary values-based principles for the responsible stewardship of trustworthy AI:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
- AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
- **There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.**
- AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.
- Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

Of note is the work done by the Defense Advanced Research Projects Agency[3] (DARPA), which has been focusing on eXplainable AI (XAI) for many years, funding research to improve the explainability of advanced AI methods, distinguish the different types of explainability and develop metrics for the trade-off between accuracy and explainability. However, best practice is still evolving around exactly how to interpret AI and AI-driven decisions, who should have access to what type of information and what it means for the financial services sector in practical terms.

Perhaps the clearest and strongest addition is the EU's General Data Protection Regulation (GDPR) which gives individuals various types of 'right of explanation' and asks for greater algorithmic transparency and auditability – seen by some as the 'gold standard' for businesses.[4] Currently the most comprehensive guidance on how to interpret and explain AI decisions is from the UK Information Commissioner's Office, prepared in response to the request of the UK government's AI Sector Deal. This guidance is 'horizontal', not specific to financial services, but provides a detailed set of considerations for firms (see section 3c).

In the UK, the FCA together with the Alan Turing Institute, is working to gain a better understanding of the explainability challenges that arise when applying AI in the financial services sector. They hope to explore the practical application of the FCA's initial framework with industry and civil society stakeholders and to shed more light on why AI transparency matters for finance.

This paper seeks to advance the thinking on how financial services firms can implement a framework that supports explainable AI – thus building trust among consumers, shareholders and other stakeholders, and helping ensure compliance with emerging regulatory and ethical norms.

3. https://www.darpa.mil/program/explainable-artificial-intelligence

4. https://www.law.ox.ac.uk/business-law-blog/blog/2018/05/rethinking-explainable-machines-next-chapter-gdprs-right-explanation

3. NAVIGATING THE CORE CHALLENGES OF AI EXPLAINABILITY

a. An introduction to explainability

FOCUSING ON TRUSTED AI WILL HELP FIRMS GET EXPLAINABILITY RIGHT

AI now impacts nearly all aspects of customers' lives, with the emergence of technologies such as smart devices in homes and offices, chat bots participating in conversational commerce, and more recently, the introduction of track and trace applications.

Every incremental use of AI, before being widely adopted, first challenges whether the particular AI-based system or technology, including the context in which it is used, can be trusted.

In recent years, stakeholder groups, including the general public, media and government at various levels, have been expressing concerns with increasing regularity in relation to the transparency and impact of AI decision-making on society. Maintaining and increasing societal trust in business remains important and we believe financial services firms should take the opportunity to get ahead in the 'trust' equation before the use of AI-based systems increases significantly.

Explainability is a fundamental component of achieving this, and implementing robust AI frameworks and governance models is key to support and justify the development and monitoring of algorithmic models throughout their lifecycles.

The assumptions and hypotheses of each application's purpose and the context in which the algorithm is deployed play a crucial role in defining the level of transparency and explainability needed for the end recipient.

Although a key consideration, firms should not focus solely on explainability. Firms should start with the overarching goal of ensuring that their AI systems are trustworthy. Beyond explainability and transparency, this includes ensuring systems are robust, stable, secure and properly governed, that data is protected and that adherence to AI (ethical) principles can be demonstrated.

Focusing on the overarching goal of trustworthy and ethical AI will require firms to keep on top of all of these considerations and will ultimately steer firms towards developing models and systems that are more transparent and explainable.

BOX 3 – AI TRANSPARENCY AND AI EXPLAINABILITY

AI transparency and AI explainability are closely linked concepts but definitions can vary.

Broadly speaking, 'transparency' relates to designing and building AI systems in such a way that there can be effective oversight. This involves the firm being able to describe the data and features used, the mechanisms by which outputs are generated and how decisions are made. 'Transparency' can also be used in a general sense to refer to being open and clear about AI systems.

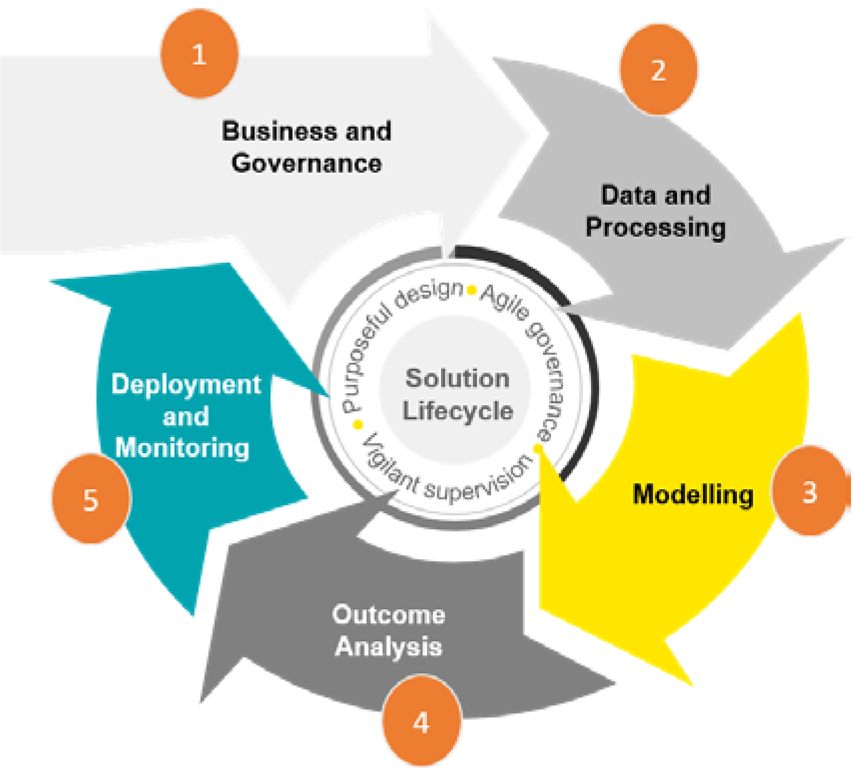
'Explainability' can also be interpreted in different ways but broadly refers to taking the technical elements of the AI system and providing a 'translated' explanation that is comprehensible to human beings. In particular, this relates to being able to explain why a specific decision was taken or why a specific output was reached.

By providing greater clarity about how AI systems operate, transparency and explainability can help firms more easily satisfy other criteria for trustworthy AI such as fairness, managing bias and ensuring accountability.

For further information, see for example:

- Fjeld et al, pages 41-43
- Independent High-Level Expert Group on Artificial Intelligence, page 18
- Information Commissioner's Office, pages 11-13

The diagram below sets out how trust in AI and thus increased transparency and explainability can be achieved through the lifecycle of AI solutions. Five components are considered; each of them is characterised by a set of attributes and corresponding testing procedures which can be tailored to different industry sectors and use cases.



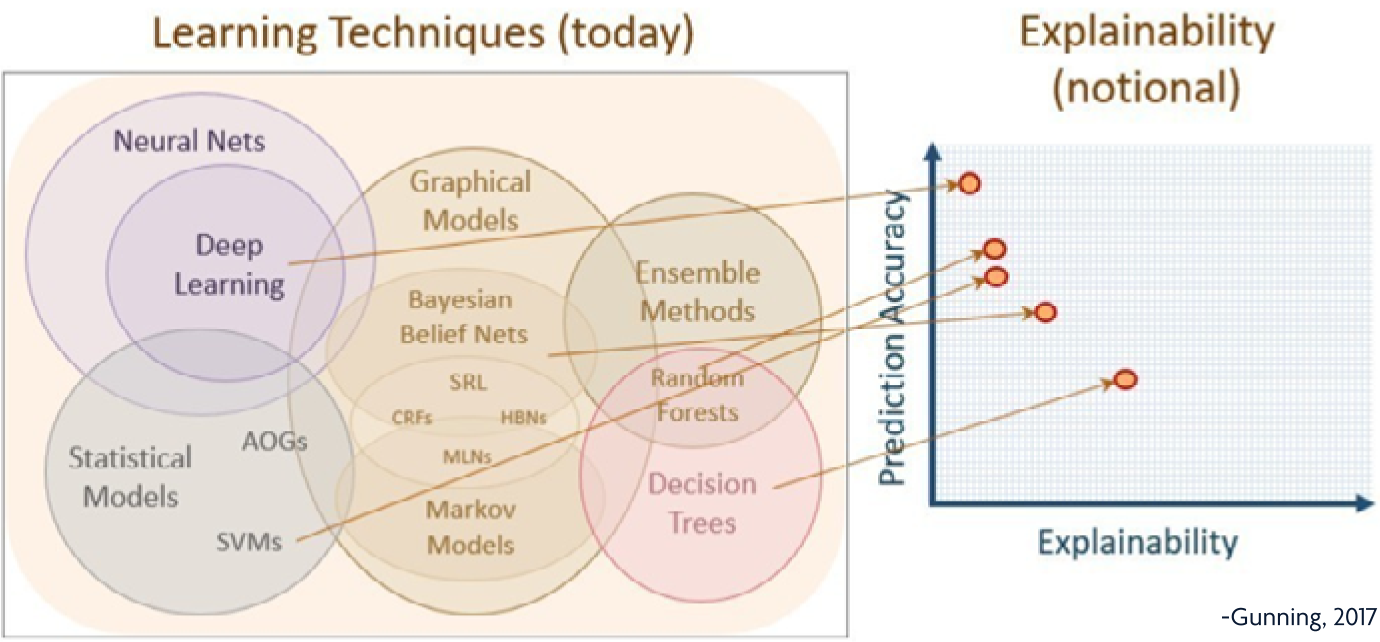
-EY Trusted Solution Framework, 2019

1. Business and governance: ensure business purpose, governance and stakeholder engagement are properly identified and aligned.
2. Data and processing: review data sourcing, profiling, processing and volumes, as well as data quality and ethical issues such as diversity in data sets and bias risks.
3. Modelling: asses approach and ensure models are fit for purpose, explainable, reproducible and robust, with supporting evidence.
4. Outcome analysis: confirm outcomes achieve desired level of precision and consistency, and are aligned with ethical, lawful, and fair design criteria.
5. Deployment and monitoring: ensure solution is scalable and deployable with the right tech infrastructure, and continuously monitored.

How and when to explain AI-based decisions and what constitutes an acceptable explanation is still subject to further careful thought and development. We are seeing increasing momentum, particularly through the work of regulators and auditors, who are becoming more adept at developing their own capabilities for auditing 'black boxes'. Firms should consider explainability as a means to promote trust with customers, regulators, auditors and other stakeholders.

b. The accuracy-explainability trade-off

It is important to understand the trade-offs when developing an AI-based system. In particular, more complex and powerful algorithms are often more accurate (for example, artificial neural networks). But that comes with a trade-off as they are usually less explainable and transparent, and potentially more costly to monitor.



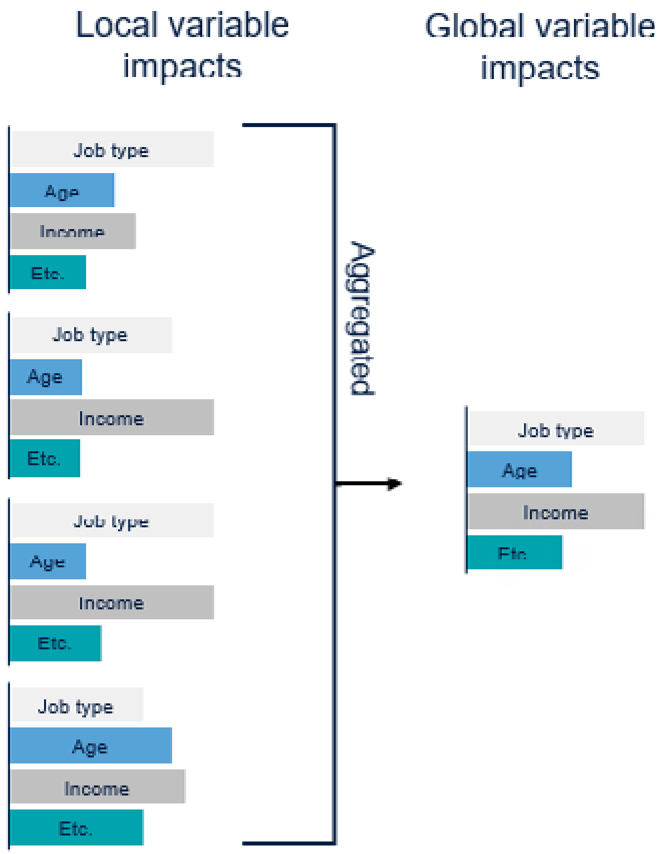
-Gunning, 2017

There is no definitive answer as yet as to how to resolve this trade-off. Regulators, government, companies and individuals will need to think through use cases carefully, considering the context and purpose for which an AI - based system is being developed. DARPA and other organisations are currently conducting research to try to better understand this trade-off.

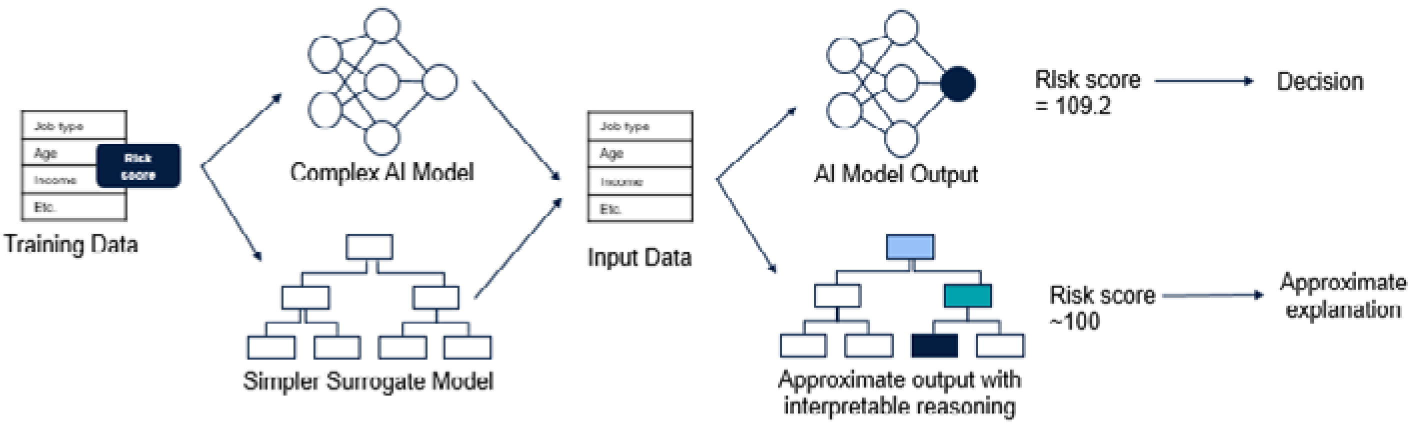
Some algorithms produce models that are more inherently explainable than others (e.g. linear regression, decision trees, rules-based approaches). However, most deep learning-based models are created with very complex artificial neural network architectures that are not inherently explainable. Nonetheless, there are methods that can be used either directly or indirectly to extract information about the rationale of the outcomes.

Numerous techniques exist that can help render otherwise opaque models more explainable. Which is most appropriate will depend on the model, data and context – but options include the following:

- Visualisation – plotting the relationship between features and the prediction outcomes helps to identify hidden patterns and better address the underlying relationship between the two variables.
- Counterfactual explanations – these involve running a model to generate an output, then changing the feature values (a new 'counterfactual' scenario) and analysing the model's new predictions. By doing a comparison between the two scenarios, a firm can infer to what extent input features change the prediction in a relevant way and identify the smallest change to feature values that impact the outputs of the model. For example, "if this customer earned £X more a year, they would get a credit score of Y".
- Surrogate models for more complex models such as neural networks or reinforcement learning – surrogate models are simplified, interpretable models that are used to explain individual predictions of 'black box' machine learning models. This is done using techniques such as dimensionality reduction (e.g. through principal component analysis) to produce a less complex model that reproduces the behaviour for a specific prediction or decision. This simplified model will be less accurate (or only accurate for a limited range of inputs) but can be more easily interpreted and explained.
- Local and global feature importance – this approach allows users to focus on the contribution of individual features for a specific prediction (local importance) and contrast it with the contribution of that feature when taking all predictions into account (global importance).



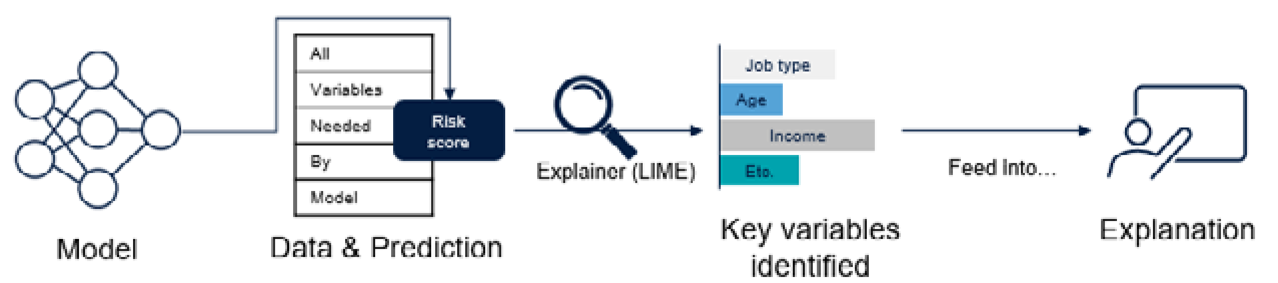
Use of a surrogate model



BOX 4 – LOCAL AND GLOBAL FEATURE IMPORTANCE

A common example of a technique to illuminate the local importance of different features is the Local Interpretable Model-Agnostic Explanation or 'LIME'. This is an approach where a firm identifies the local importance of input features by making small changes to them and observing how this impacts the model's outputs. This investigation into the relationship between specific input features and outputs can give an indirect view of which input features are having the largest effect on a specific output. This can be an important part of an explanation of a specific output, such as a specific risk score.

Shapley Additive Explanations ('SHAP') are a technique for identifying globally important features. It uses Shapley values from cooperative game theory to establish the average importance of input features for every possible combination, and to help identify the features that have the largest impact on the model's outputs in general (as opposed to in relation to a specific prediction). This global importance can, for example, inform generic explanations for all customers of how a model operates.



-Ribeiro, et al, 2016

c. ICO guidance – the most comprehensive regulatory view

i. OVERVIEW

As part of their mission to support UK organisations to be leaders in the use of AI, while being compliant with data protection laws, the ICO and the Alan Turing Institute initiated "Project ExplAIn".[5]

The guidance developed by this project and published in May 2020 does not impose new regulatory requirements and is not a statutory code of compliance. Rather, it aims to help firms understand good practice for explaining AI, as well as highlighting relevant GDPR requirements. A core premise of the guidance is that being able to explain AI effectively will help firms enhance consumer trust and more readily comply with existing regulation when implementing AI models.

In particular, the ICO guidance was designed to help firms comply with the following regulatory requirements:

- GDPR articles 13 and 14 (the right to be informed), 15 (right of access), 21 (the right to object), 22 (rights related to automated decision-making, including profiling) and 35 (data protection impact assessments).[6]
- Equality Act 2010, which requires that organisations ensure that AI systems in their decision-making process do not result in discrimination.[7]

The guidance contains three parts, addressing:

- The basics of explaining AI – for a broad audience, including data protection officers and compliance teams, but also relevant for technical teams and senior management. This section most notably sets out:
  - Key definitions and the legal framework, particularly GDPR provisions.
  - Types of AI explanation identified by the ICO.
  - Important contextual factors to consider, such as the domain, nature of the data, audience of the explanation and likely impact of the AI decision.
- Explaining AI in practice – aimed primarily at technical teams, but also relevant for data protection officers and compliance teams. This part sets out six tasks for firms, starting with the choice of what types of explanation to provide, through to considering how to deliver the explanation to end-users.
- What explaining AI means for an organisation – aimed primarily at senior executives, but also relevant for data protection officers, compliance and technical teams. This part provides suggestions for setting organisational roles and responsibilities, preparing policies and procedures, and ensuring comprehensive documentation.

Recommended process for ensuring effective explanations

The guidance suggests six distinct explanation types that could be appropriate (see box 5). For each type of explanation the firm determines is necessary, it should consider what needs to be explained about the process by which decisions are made, and what needs to be explained about the specific output (decision). Which explanations are chosen and what is included for each should be tailored to the needs of the stakeholder, the use case and the context in which the AI is being deployed.

5. See Project ExplAIn Interim Report: https://ico.org.uk/about-the-ico/research-and-reports/project-explain-interim-report/
Final guidance: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/

6. After the end of the Brexit transition period, these requirements will be incorporated into UK domestic law.

7. A particular challenge for firms is that these two pieces of legislation both create types of higher risk data that require particular care, but these do not fully align.

BOX 5 – EXPLANATION TYPES

1. Rationale explanation: a non-technical description of how the system arrived at its outcome.
   - Helps individuals challenge decisions or adapt their behaviour to get a different decision next time.
2. Responsibility explanation: who is involved in the development, management and implementation and who to contact for human review.
   - Helps individuals know how to challenge a decision.
3. Data explanation: what data has been used and how, including testing and training.
   - Helps individuals challenge a decision if data was not appropriate to them.
4. Fairness explanation: steps taken across the design and implementation of an AI system to ensure its decisions are unbiased and fair.
   - Helps reassure individuals about the steps taken to achieve fairness.
5. Safety and performance explanation: design and implementation steps to maximise accuracy, reliability, security and robustness.
   - Serves to inform and reassure individuals about the resilience of the system.
6. Impact explanation: impact that the use of an AI system and its decisions may have on individuals.
   - Useful in advance of a decision to help individuals decide whether a service is right for them.

The guidance also provides a process that can help the firm deliver meaningful explanations for decisions that are made, or assisted, by AI. This process is broken down into six tasks ranging from assessment of the type of explanation that stakeholders will be looking for, to the necessary information gathering and delivery method (see box 6). The focus is on considering explainability needs through all stages of AI system design.

BOX 6 – SIX TASKS FOR BUILDING EXPLAINABLE AI

1. Select priority explanation types (see box 5) by considering the domain, use case and impact on the individual.
   - Note that for many applications there will be more than one explanation type that may be important to stakeholders.
2. Collect and pre-process your data, while considering the explanations you will need to provide.
3. Build the system in a way that allows relevant information to be extracted.
   - Firms should choose a model providing a level of interpretability that is suitable to the use case (Annex 2 of the guidance suggests tools for 'black box' models).
4. Translate the technical explanation into easily understandable language.
   - Consider using graphics, media, tables or other tools.
5. Prepare 'implementer' staff to deploy the AI system.
   - Consider their understanding and capacity to convey explanations.
6. Consider how to build and present explanations.
   - Consider the domain, impact, data, urgency and audience to determine what information individuals needs and how to present it clearly, for example in 'layers' so as to avoid information overload.

Additional key considerations for firms

When considering how best to implement the recommendations provided by the ICO's guidance, it is important to approach this exercise from the perspective of the intended audience of the explanation. Customers will have other priorities from regulators and providing them with too much overly technical information is likely to result in confusion that might even reduce their trust.

Context is key – firms should ensure that business processes and governance are in place to trace back to the purpose of the AI system. Without context, the model and its outcomes will be difficult to understand and interpret accurately.

Although not the focus of the guidance, firms could refer to it to help inform communications with business customers or regulators.

Whichever audience is receiving the explanation, firms will need to take into account other sector-specific obligations. For example, model risk management rules, FCA requirements for explaining lending decisions and investment advice introduced to protect borrowers and investors – covering not just conventional credit institutions and investment firms but also newer business models such as the peer to peer sector.

Beyond regulatory obligations or considerations around the customer trust relationship, the ICO's guidance on explaining AI decisions can also provide a useful framework for improving existing internal policy around documenting automated decision making processes (especially the 'process' elements of explanations covered in Part 2 of the guidance).

As a separate point, when considering the purchase of AI applications from third-party suppliers, the ICO guidance may provide useful suggestions for the types of explanations and documents to consider as part of the due diligence in the procurement process. This guidance might further be augmented by the UK government's 'Guidelines for AI procurement',[8] which were co-developed by the Office for AI and the World Economic Forum's Centre for the Fourth Industrial Revolution, among others. While these guidelines were developed for public sector procurement, and thus are not mandatory for the private sector, they still provide a useful set of considerations that can help guide firms. Among the considerations are recommendations regarding assessment of data requirements, deployment risks, governance considerations, implications of 'black box' algorithms and potential for vendor lock-in, as well as the need to address technical and ethical limitations and lifecycle management.

Limits of explainability of AI

There are still known limits to explainability and transparency, even with existing techniques and approaches in place. More research is needed with regards to how best to design and manage AI systems, particularly their growing use and ability to outperform human decision-making.

8. https://www.gov.uk/government/publications/guidelines-for-ai-procurement/guidelines-for-ai-procurement

ii. PRACTICALLY APPLYING THE RECENT ICO GUIDANCE TO COMPLIANCE

Determining what is in scope

First, review the guidance and decide what is in scope for explainability in your organisation. Consider models used for conduct as well as operational and prudential purposes and consider which systems 'make decisions'. Think about credit decisioning as well as algorithms used to send alerts to customers to promote new deals, to alert customers who are about to go into overdraft, and those who are in potential or actual financial difficulty.

Consider areas of heightened interest to regulators where customers are generally more vulnerable, for example whether dialler systems for contacting customers might be deemed to be unduly pressurising customers. AI-specific risk criteria from regulators could also become relevant, for example under proposals for stricter regulation of 'high risk' AI under the European Commission's AI white paper.

> Consider applying a wider scope so AI-based systems can be prioritised and deliberately 'parked' rather than excluded from the outset. This is particularly important where an AI-based system may initially be considered to be low impact, but if its scope or population impacted is increased, it would become higher impact. The reason for a decision to 'park' a model should be documented in case it is challenged at a future time. It may help to 'tier' models to more easily see when they might need to be prioritised.

Once firms have decided which types of models are in scope for explainability, consideration should be given to creating an inventory of those models, keeping this up to date and determining which of these models should be prioritised for explanation. Example principles for prioritising models might include:

- Type of outcome, with outcomes that will impact the customer's access to finance or impact the type, scope and size of product they are entitled to being higher priority, especially where a fully automated 'significant decision' is being made under GDPR Article 22.
- Number of customers impacted, with a greater number of customers impacted being higher priority. This parameter should consider model, customer and business growth plans and whether any of these are likely to significantly increase the number of customers impacted by the model.
- The extent to which documentation and subject matter resources are available to help explain the model, with models with limited existing documentation (e.g. due to the age of the model) to support explainability being higher priority.
- Functions or departments where there is reliance on AI-based systems and either a greater potential for the customer to dispute or challenge an outcome, or a high number of complaints. Consider also the extent of reliance on AI.
- Areas where unfair bias could be deemed an issue, where different types of customer may receive different outcomes, and where there could be a higher risk of protected characteristics having an unintended influence on outputs (e.g. where certain data points could 'proxy' for protected characteristics). This parameter tends to be important for financial models, for example those which determine credit availability for a customer. Often, model inputs are chosen to achieve prudentially sound outcomes based on risk factors, in line with the firm's risk appetite. Particular care should be taken to ensure those decisions do not indicate unnecessary bias, and that outcomes can be explained.

The principles which firms use to determine priority for model explainability will need to evolve as more models are deployed and use cases expand.

Explanations may already exist and firms should review them in light of the ICO guidance, again systematically determining priority.

Good AI governance and clear accountability

Having determined what is in scope, firms should collate their inventory of models and put in place a process to keep it up to date, leveraging GDPR inventories where possible. Policies, processes and procedures for maintaining good governance over AI-based systems, such as guidance on explainability, should be widely promulgated within the organisation, ensuring that all employees who engage with setting up models are fully aware of their importance for maintaining customer trust and regulatory compliance.

Using the example principles set out above, firms can develop risk and impact scoring methodologies to help them assess which models should be prioritised as well as thresholds and triggers for determining when a model might become a priority in future. Ensure that pilots and proofs of concept are covered, with lighter touch governance where appropriate.

Model review frequency should also be considered (both internal and external). For example, this should be dependent on the relative priority of the model, the last date it was reviewed for explainability and the extent of change the model (or business in which it operates) may be subject to, etc.

When determining who is on model governance committees, consideration should be given to including representatives of stakeholders who might need an explanation or be required to provide an explanation (for example customer-facing staff) as well as those developing the models. The adequacy of explanations will need to be approved by the model adequacy committee, another suitable committee or a suitable individual. Those deemed to be 'accountable' for the explainability of a model should be on the governance committee.

Deciding who is 'accountable' – an accountable individual will need to have sufficient authority and knowledge to oversee and direct the process by which the model is developed or operated, including the use of third-party systems or open source elements. Thought needs to be given to the interplay with existing senior manager and certification regimes, where a firm has to comply with these. For example, if a model is built in second line but it impacts customers in a line of business, who should be accountable for the explainability of the model and how will the firm limit duplication of effort?

Those who are accountable will need to have access to monitor the performance of the models, including the ability to investigate anomalies and interrogate the models through independent testing and validation. Accessibility for auditors and regulators will also need to be considered – will access be via documentation or directly into the AI-based system? Likewise, firms should consider how testing will be made possible for auditors and regulators.

Where an external developer is used, a common understanding of the roles and obligations of both the developer and the implementing firm will be needed to ensure that both parties have the information they need to generate any explanations relevant to their role. These could be for internal audiences (e.g. control or compliance) or external audiences. This should include ensuring that compliance with regulatory requirements can be demonstrated by both firms.

Determining what type of explanation is needed: know your customer

The ICO guidance is focused on explanations to be provided to end-users who are individuals, mainly customers but also potentially employees, depending on the use case (it is not designed with business-to-business use cases in mind). Such outward-facing explanations of AI systems and AI decisions intended for customers will need to be consumer friendly and focused on the information they need most. Customers will often be interested in understanding why a particular decision about them was reached, particularly if they are not happy with the outcome. The ability to explain decisions to customers in a way they can understand is an important contributor towards securing their trust.

Firms should be mindful of how existing disclosure requirements may interplay with the ICO's guidance. Good explanations will need to take into account the best timing for the explanation and whether it should be 'layered' with other customer information needs.

Although not the focus of the ICO guidance, AI explanations can also be focused internally. Internal explanations can be important for control and performance monitoring purposes, to ensure that the model is behaving correctly. This will require standard methodologies and tools which can show how the model compares with performance, ethical, legal, and regulatory criteria. A clear view is needed of what 'fair' outcomes will look like, taking into account the assumptions and hypotheses of the specific application, purpose and context, and an approach for the identification and mitigation of unfair bias.

d. What it means for financial services – exploring potential use cases

START WITH THE CONTEXT AND THE 'WHY', THEN CONSIDER THE 'WHAT' AND THE 'HOW'

This section reviews some financial services examples to explore what types of explainability may be needed and how to communicate effectively to the end consumer. The case studies also consider what approaches to explainability, transparency and information are needed internally. The first use case looks at specific AI-enabled products and what kinds of explanations might need to be produced, particularly for customers and consumers.

The following two tech use cases focus more on the 'behind the scenes' technical considerations, which are key to building effective explanations intended both for internal control purposes and as inputs for explanations provided outwards to consumers. These focus on models that are more complex and less explainable, given the shift towards these in financial services (and other sectors).

Use cases

BOX 7 – BUSINESS CASE STUDY: CREDIT DECISIONS AND ALGORITHMIC BIAS

Problem: Decisions to provide a loan, issue a credit card or write a mortgage could be made using AI-based systems. However, we have already seen incidences globally of credit providers being accused of biased and unfair decisions due to:

- Offering a different credit limit to men and women who appear to have the same financial fundamentals.
- Basing creditworthiness decisions on limited historical information available for customers who are recent immigrants.
- Basing a creditworthiness prediction based on incomplete or erroneous information.
- Making creditworthiness predictions based on simple criteria such as browser settings and whether the applicant uses a Mac or PC.

It is possible for algorithms to inadvertently discriminate on the basis of education, income or even protected characteristics such as age or gender, even when that variable is not explicitly included as a factor for the decision process. Effective explanations can help prevent this risk from arising.

Recipients for explanations: The business, regulatory compliance teams and the end consumer.

Key considerations:

- Customers will often want to know why they have been declined credit or why they have received a specific credit limit. They might want to increase their access to credit in the future, and information about the reason for the lender's decision could help them improve their creditworthiness, for example by paying down existing debt. At the same time, firms will need to balance against the risk of customers trying to 'game' the system illegitimately.

- Access to credit is a sensitive issue, so when customers disagree with their credit limits, they might suspect that a biased decision has been made. An explanation of how the firm ensures decisions are fair can help in the event of complaints.
- GDPR highlights that providing an explanation of automated decisions can be an important customer protection, while under the Standards of Lending Practice a firm should provide the principle reason for a 'decline' decision.
- Internally, given the reputational, customer and regulatory risks associated with making unfairly biased (or even discriminatory) lending decisions, the firm will want to be sure that it can 'explain' to itself why certain decisions have been made. This will help the firm be sure that it is making credit decisions on a sound basis and is meeting its regulatory obligations to be fair, lend responsibly, and not to discriminate.
- The use case also needs to be built in line with the wider business case, including factors such as risk appetite and profitability.

Key types of explanation: rationale explanation; responsibility explanation; data explanation and fairness explanation.

Preparing and delivering explanations:

- The firm should consider choosing a more interpretable model (eg. linear regression or random trees). However, trade-offs should be evaluated, depending on the accuracy needed.
- Ensure data and models are fair and traceable and align with the use case purpose. This process must be documented. This will help the firm reassure itself that it is making transparent and unbiased decisions.
- If using unconventional data, such as digital footprint (device type, browser settings, email host, etc). Or a large, unstructured data set such as social media data, extra care will be needed with the internal data and fairness explanations. Firms will need to ensure that decisions are well understood, with robust processes in place to address risks of unfair bias. Identify variables and any potential bias applying to certain classes of customer, paying particular attention to the risk that certain data could 'proxy' for protected characteristics.
- An internal responsibility explanation will help ensure the firm knows where to follow up in the event of a complaint.
- In terms of an explanation to the customer, the lender could start by providing a basic rationale explanation in the first instance, setting out the main factor(s) that impacted the decision and making use of the Standards of Lending Practice or other industry standards, where applicable. Using a 'counterfactual' explanation technique could assist. It may also help to provide a light-touch responsibility explanation so the customer knows where to take any concerns or complaints.
- If the lender will be relying on unconventional data or large, unstructured datasets, an up-front data explanation could be helpful. Customers are more likely to complain if they learn, after the fact, that surprising data has been used as an input.
- The lender might wish to set up a process to produce a more detailed rationale explanation if asked, for example a process-focused explanation of why applicants are denied credit or get different credit and mortgage rates. A more comprehensive 'on request' data explanation and a fairness explanation could also be useful to help reassure the customer that a fair, unbiased decision has been taken. A careful balance would be needed in any customer-facing explanations, to ensure that helpful information is provided without showing customers how to game the system illegitimately, or unnecessarily taking up customer time.

BOX 8 – TECHNICAL CASE STUDY: PERFORMANCE MONITORING FOR REINFORCEMENT LEARNING MODELS IN PRODUCTION

Problem:

Reinforcement learning (RL) is a class of AI models that are inspired by behavioural psychology used by artificial agents to learn autonomously via interactions with their environment. RL can be used, for example, in a lender's debt collection use case for strategy optimisation and resource allocation, being able to customise the collection approach for each client in real-time.

Firms will need to achieve adequate explainability for internal assurance purposes and to provide sufficient explanations to regulators and customers, as required. An open issue in RL is the lack of visibility to understand the decisions taken by a trained agent during the learning process, making it difficult to understand how inputs get transformed into outputs. Nor is this readily explainable. Addressing this challenge requires a mixture of technical and organisational measures.

Firms will also need to be mindful of their FCA obligations to engage effectively with customers and agree an affordable approach to repayment.

Why explanation is needed:

Collection models must exhibit a degree of transparency for business users to understand whether changes in inputs lead to changes in outputs that have explanations consistent with the insights of human domain expertise. A lack of explainability can undermine validation and production because it makes it difficult to understand whether models are successfully meeting testing objectives and are fit for purpose.

Firms should also ensure that their decision-making process to decide on a choice of model is transparent and auditable.

The need for explainability is motivated by the need for trust, interaction and transparency between the end-user and the RL system. Thus, before deploying an RL model in production, the model developer and validator should assess whether an appropriate explainability framework is required. The choice of framework should be based on the intended business outcome, the type of end-user and the established override mechanisms. Human oversight, and the possibility to challenge and subsequently override a model's outcome, is of high importance. In some cases, machine learning algorithms can pick up certain features humans would not, leading to different outcomes (and vice versa). Exploring those outcomes and establishing override mechanisms that rely on individual judgement indicates that a model developer does not blindly follow the machine learning (ML) model. Such practices enhance explainability and demonstrate the application of professional reasoning.

How to deliver an explanation:

As outlined in more detail above, the key frameworks for enhancing AI model explainability can be grouped into:

- visualisation
- local and global feature importance
- surrogate model

The firm's choice will depend on the algorithm and application. For instance, understanding local importance is relevant in supervised learning models. Importance can be evaluated 'locally', where the firm assesses the effect that the attributes of an individual observation have on the model's prediction for that observation. Visual inspection is more appropriate for unsupervised learning models and reinforcement learning models, while surrogate models allow the firm to simplify overly complex models, such as deep neural network, including Deep Q-Networks (DQN) or Natural Language Processing (NLP).

Key steps to build for ongoing explainability:

1. Create an integrated reinforcement learning model performance monitoring framework for evaluation and approval of the re-trained models, monthly performance monitoring and trigger reviews, and tracking and reporting of issues covering all controls.
2. Setting thresholds for reviews should consider the top down risk appetite from the board and senior management to make sure there is objectivity and consistency across models.
3. Offline performance test includes swap-in swap-out analysis, backtesting (test the DQN), with action recommendations on top and bottom historical percentiles for the one-step return, and DQN long-term return prediction checks.
4. For the online learning performance monitoring, establish criteria to compare the champion and challenger model (e.g. traffic light approach) and set up the approval process required to swap the champion and challenger.
5. In the online model retraining, set the exploration rate in conjunction with business risk tolerance. The exploration rate should decrease/decay over time. Develop overlay to exclude high impact accounts (e.g. high estimated credit loss) from exploration. Isolate model retraining (development) from model prediction (production).

BOX 9 – TECHNICAL CASE STUDY: USING EXPLAINABLE AI FOR ML VALIDATION TO ADDRESS ALGORITHMIC BIAS

Explainability from the business perspective:

The responsibility for de-risking and enabling safe innovation in AI/ML spans multiple parts of the organisation. Both model developers and validators should practice model explainability methods as a part of controlling for algorithmic bias and ensuring fair outputs.

Beyond technical limitations, the business environment may impose additional constraints on model validation. It is important for every validation to address not just the model methodological aspect (how the model behaves) but also the business use aspect (the considerations that are specific to the use case).

From a business perspective, **model transparency** and **model fairness** are the most important aspects of a sound and trustworthy AI/ML model.

- **Model transparency** can be described as the degree to which a human can understand the decision framework of the prediction, or the degree to which a human can consistently interpret and predict the model's result. A key element of transparency is understanding how important each element of input data is to the accuracy of the prediction or other output. This 'feature importance' can be determined, for example, through permutation of single feature values; varying them one by one to see how much they affect the output. The results should be reviewed and discussed with the business users. Any deviation from the business expectations should be analysed and justified.

  Sensitivity analysis is used to understand the effect of a set of features on some target variable under certain specific conditions. This can be performed to avoid counter-intuitive model behaviours under different market conditions.

- The second aspect, **model fairness**, means ensuring that the model does not lead to outcomes which negatively impact a subset of the population in an unjustified way. For example, unfairness can arise if predictions are based on data that reflects institutional, societal or historical bias (e.g. in relation to gender or race). There are two types of fairness: individual fairness and group fairness. 'Individual fairness' requires similar individuals to be treated similarly. 'Group fairness' attempts to ensure that members of all protected groups receive a fair share of beneficial outcomes. Firms should consider both when developing a model and reviewing its outputs. An assessment of fairness should begin with a clear and documented statement of the fairness definition and its relevance to the underlying business application. For example:
  - Demographic parity, where the same fraction of each demographic group gets a 'positive' outcome.
  - Equalised odds, where the 'true positive rate' and 'false positive rate' are the same for each demographic group.
- Consideration should be given to whether a formal non-discrimination criterion (such as an equal ratio of positive to negative outcomes for males and for females) is necessary in the objective function and associated transformation logic should be deployed.

- Several types of criteria such as independence criteria, separation criteria, and sufficiency criteria can be used. The appropriate criterion to select will depend upon how fairness is interpreted in the context of the business decision, and on the data available to analyse different demographics. If applicable, the boundaries of permissible 'positive action' under equalities law should be considered.

If a firm achieves a high standard of model transparency, it can more effectively provide 'rationale explanations' to customers or other end-users. It can also more readily confirm that it is satisfying its fairness principle and non-discrimination criterion. This will help firms prepare robust 'fairness explanations' when needed.

4. FUTURE OF REGULATION AND FINANCIAL SERVICES

THE REGULATORY HORIZON WITH REGARDS TO XAI (AND ITS LIMITATIONS)

Where the regulatory focus will be in the near future

As use of AI and ML systems increases in all industries, government and regulators across the globe will continue to seek to understand the efficiencies they can bring to the industry and the protections needed to mitigate risk to customers. This could take the shape of new guidance but will also involve an assessment of how existing regulation and legislation applies to the use of AI and ML, and whether further interpretations are needed. As seen in the recent European Commission data strategy consultation, authorities are also exploring how to encourage the use of data to drive benefits through AI and ML (including public policy benefits), so the regulatory and legislative focus is likely to see a combination of encouragement for AI uptake and efforts to mitigate the risks it may bring.

Explainable AI will be a key feature of this focus. It already plays a prominent role in many of the Principles for Ethical AI that have been adopted (see section 2b above). Regulators are now focusing on how to translate these principles into practice. What will be the requirements for compliance with this principle? Where possible, international alignment of principles will be an important way to reduce obstacles to firms operating across jurisdictions and protect customers across borders.

Globally, a working group of the International Organization of Securities Commissions (IOSCO) is considering AL and ML. It aims to establish a broad sector-wide consensus around the risks that need to be managed and how the costs and benefits of that burden should be carried. It will also map the areas where industry needs regulatory clarity to invest further in AI and what obstacles to avoid in terms of regulation within single jurisdictions and internationally.

Some of the important contributions shaping the IOSCO consultation are from the Monetary Authority of Singapore's Veritas initiative on 'fairness metrics to aid responsible AI adoption in financial services'[9] and the Hong Kong Monetary Authority's 'High-level Principles on Artificial Intelligence'[10], which include principles around accountability of boards and senior management for AI-related outcomes.

In Europe, the European Commission is exploring the adequacy of current regulatory frameworks for dealing with the challenges posed by the growing use of AI. The intention is to position the EU as a leader in 'Trustworthy AI'. After a year-long review by the High-Level Expert Group on AI, which published two reports[11] in 2019, the Commission in February 2020 published a white paper on AI.[12]

9. https://www.mas.gov.sg/news/media-releases/2020/fairness-metrics-to-aid-responsible-ai-adoption-in-financial-services

10. https://www.hkma.gov.hk/media/eng/doc/key-information/guidelines-and-circular/2019/20191101e1.pdf

11. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top ; and https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

12. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Regulatory fields that have been identified as particularly important to fundamental rights in relation to AI include financial services, migration and responsibilities of online intermediaries. The white paper indicates that the European Commission intends to pursue a risk-based approach to the regulation of AI, with sector specific considerations as one of the dimensions in the risk assessment.

Among the questions being asked by regulators is whether to regulate the technology or its application for specific use cases (e.g. demand explainability for ML models, regardless of which domain they are used for, or explainability for all medical diagnostics tools, regardless of the technologies used). While not having expressed a clear answer to this question, the preparatory work indicates a compromise position. The high level principles from advisory bodies such as the European Commission's High-Level Expert Group on AI have been formulated mostly as cross-sector, technology-oriented recommendations, whereas the emphasis on updating existing sector specific regulation suggests an enforcement strategy based on use cases.

In the UK, the Alan Turing Institute together with the FCA is working on guidance regarding AI transparency.[13] It is anticipated that this guidance will include an acknowledgement that the financial services sector has already done much work to address concerns around clarity of models and their explanations to regulators and users, following the financial crisis in 2008. Based on the supplementary measures to the Basel II risk-based framework that were passed in response to the 2008 crisis, firms have been required to meet requirements for safeguards that include transparency and explainability dimensions. Firms will need to continue to ensure that the AI models and data they use are appropriate to the context and purpose, and that these decisions are well documented.

This work does not take place in a vacuum, and it will be important to consider horizontal sector-agnostic ways to protect customers and mitigate risk, recognising that in a digital economy a customer interacts with AI systems in all sectors. 'Supervision by activity' is therefore an important lens through which to consider customer protections regarding AI.

13. https://www.turing.ac.uk/news/ai-transparency-financial-services

5. CONCLUSION

Explainability of AI is used in many different contexts, including to better engage with businesses and end consumers. The context and purpose of the AI system is key to determine what type of explanation is required and what information exactly should be shared.

It should be a firm's objective not only to provide AI explanations but also to ensure that their AI can be trusted, and is fair and accurate.

So far, the regulatory focus is geared towards providing explanations for individuals. However, when engaging with business customers, or if providing an explanation to a regulator or professional body, the ICO guidance can still be a useful resource.

Whichever audience is receiving the explanation, firms will need to take into account other sector-specific obligations and the trade-offs between accuracy and model complexity. Most models currently used are more or less explainable – however, given artificial neural networks and other deep learning methods will likely become more prominent in the financial services sector over the next years, it is important for firms to have the right AI strategies, reporting governance and frameworks in place in order to stay compliant and trustworthy.

It is also important to keep in mind that this is an evolving domain. There are still known limits to explainability and transparency, though helpful techniques and approaches exist. More research is needed with regards to how best to design and manage AI systems, particularly given their growing use and ability to outperform human decision-making.

KEY TAKEAWAYS

Good governance for firms:

- The purpose and context of the AI being used should be clear, with the right data and models used and documented.
- Governance to monitor and report on AI systems should be agreed, with a clear fallback and accountability plan, in case the AI goes wrong.
- Scoring methodologies to help assess which models should be prioritised for explanations or enhanced explainability, as well as thresholds and triggers for determining when a model might become a priority in future.
- Procedures to follow when trade-offs need to be made between different AI trust attributes, e.g. explainability vs. accuracy.
- Firms should keep in mind the limits of explainability, even with techniques in place to enhance model interpretability. Explainability should be seen as a part of a wider approach to trustworthy AI.

Preparing effective explanations:

- Different users require different types of explanations in different contexts - ‘know your customer’ – explanations vary depending on purpose and recipient. Be clear on why an explanation is needed, what type of explanation is needed and which information is to be shared.
- Stakeholders within the firm should be engaged - compliance, data science, product and other teams will need to be involved to ensure technical accuracy and full consideration of context.
- Articulate limitations and challenges such as privacy, data quality, data provenance and the reliability of AI explainability. Explanations may ensure greater trust in the short term but may not be enough on their own to create systems generating trustworthy outputs; it is therefore important to consider how trust in AI will be achieved holistically.

Future considerations for regulators:

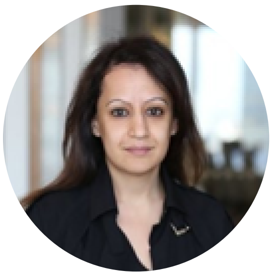
- Potential for alignment of scope of ‘data ethics’ and AI guidance across different authorities; the detail of what is needed in an explanation will vary depending on sector, service, and wider context.
- Prioritisation of existing rules and regulations for AI-specific guidance.
- Alignment of definitions and terminology.

6. REFERENCES

- Casey, Bryan; Farhangi, Ashkon and Vogl, Roland (2018) Rethinking Explainable Machines: The Next Chapter in the GDPR's 'Right to Explanation' Debate. [Available at] https://www.law.ox.ac.uk/business-law-blog/blog/2018/05/rethinking-explainable-machines-next-chapter-gdprs-right-explanation (Accessed 13 July 2020).
- Columbus, Louis (2019) Why AI Is The Future of Financial Services. [Available at] https://www.forbes.com/sites/louiscolumbus/2019/08/15/why-ai-is-the-future-of-financial-services/#7cbeb8a73847 (Accessed 13 July 2020).
- FCA (2019) Press release. [Available at] https://www.fca.org.uk/news/press-releases/fca-confirms-new-rules-p2p-platforms. (Accessed 13 July 2020).
- Fjeld, Jessica; Achten, Nele; Hilligoss, Hannah; Nagy, Adam and Srikumar, Madhulika (2020) Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI [Available at] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482## (Accessed 13 July 2020).
- Gunning, David (2017) Explainable Artificial Intelligence. [Available at] https://www.darpa.mil/attachments/XAIProgramUpdate.pdf (Accessed 14 July 2020).
- Independent High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. [Available at] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (Accessed 13 July 2020).
- Information Commissioner's Office (2020) Explaining decisions made with AI. [Available at] https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/ (Accessed 13 July 2020).
- Paudyal, Prajwal (2019) Should AI explain itself? or should we design Explainable AI so that it doesn't have to. [Available at] https://towardsdatascience.com/should-ai-explain-itself-or-should-we-design-explainable-ai-so-that-it-doesnt-have-to-90e75bb6089e (Accessed 13 July 2020).
- Ribeiro, Marco Tulio; Singh, Sameer; Guestrin, Carlos (2016) Local Interpretable Model-Agnostic Explanations (LIME): An Introduction. [Available at] https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/ (Accessed 13 July 2020).
- Royal Society (2019) Explainable AI: the basics – policy briefing. [Available at] https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf (Accessed 13 July 2020).
- Schmelzer, Ron (2019) Understanding explainable AI. [Available at] https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/#130b4c1c7c9e (Accessed 13 July 2020).
- University of Cambridge, World Economic Forum, EY and Invesco (2020) Transforming Paradigms: A Global AI in Financial Services Survey. [Available at] http://www3.weforum.org/docs/WEF_AI_in_Financial_Services_Survey.pdf (Accessed 13 July 2020).

7. CONTRIBUTORS

EY



Sajedah Karim
Partner, Financial Services RegTech and Compliance Transformation Leader, Ernst and Young LLP

Ariane Buescher
Global Client Technology AI Risk and Trusted AI Leader, Ernst and Young LLP

Ansgar Koene, PhD
Global AI Ethics and Regulatory Leader, Ernst and Young LLP

Mario Schlener
Partner, Financial Services Risk Management Leader, Ernst and Young LLP

Jason Tuo
Risk Analytics and Solution Leader, Ernst and Young LLP

Luis Pizarro, PhD
Global Client Technology AI Quality Lead and UK Client Technology AI Centre Leader, Ernst and Young LLP

UK FINANCE

Walter McCahon
Manager: Data Policy

Jonathan Middleton
Principle: Technology and Digital Policy Delivery Coordination

Oliver Nelson-Smith
Analyst: Digital, Technology and Cyber