

**College of Humanities
Master of Financial
Technology-
Fin Tech**

**Artificial Intelligence for
Financial Technology**
FNT 626

Lecture 02

**Prepared By:
Dr. Abeir Moussa**



Artificial Intelligence for Financial Technology
FNT 626- AI for Fin Tech

Lecture 02

Prepared by
Dr. Abeir Moussa

LECTURE
02

CHAPTER TWO



EXPLAINING EXPLAINABLE AI



LECTURE
02

Ch. Explaining explainable AI

- **Learning Objectives**

1. Understanding the Need for Explainable AI
2. Exploring the Principles of Explainable AI
3. Understanding the Local Interpretable Model-Agnostic Explanations (LIME)
4. Explore the applications of explainable AI in various domains
5. Stay updated on advancements in Explainable AI
6. Addressing Challenges in Explainable AI

LECTURE
01

Ch2. Explaining explainable AI

- **Chapter outline**

1.INTRODUCTION

2.WHAT IS AN EXPLANATION?

3.WHY ARE EXPLANATIONS NEEDED?

4.HOW TO MAKE EXPLANATIONS EFFECTIVE

5.CHALLENGES

6.RECOMMENDATIONS TO FINANCIAL INSTITUTIONS ON CREATING
EXPLAINABLE AI

7.CONCLUSION

LECTURE
02

Ch.2 Explaining explainable AI

1. Introduction

As companies' use of AI increased, there was a growing recognition of the importance of transparency due to the risks posed by AI models, and that the so-called 'black-box' nature of AI was a critical blocker to broader adoption of the technology.

Many techniques were (and are) published without being tested on human subjects to assess if the explanations that are generated translate to actual changes in variables like increased trust, knowledge, or efficacy. Perhaps, partially as a result of this mismatch, academic interest in explainable AI has only partially translated into more transparent commercial AI systems. For instance, in the United States (US), financial institutions must provide clear reason codes when an applicant is refused credit, whether the decision was based on a rules-based system or modern AI.

LECTURE
02

Ch.2 Explaining explainable AI

To develop, use, and supervise AI systems effectively and responsibly, stakeholders need to understand aspects of data collection, data processing, model building, model usage, and more. The second section reviews the types of stakeholders who might want explanations, and the legal obligations, risk management concerns, and business imperatives these stakeholders might have that create the need for explanations.

For instance, a risk manager might ask why a data scientist believes that a model will work well once it is deployed, and a user might ask why they should take the action of accepting an algorithm's recommendation. In fact, there are three different types of questions that the explainee might have, which correspond to three different types of reasons. Normative reasons ask why something is a good idea or the right thing to do.

2. WHAT IS AN EXPLANATION?

2.1 Explanations of Internal Reasons

2.2 Explanations of External Reasons

2.3 Explanations of Normative Reasons

2.4 Types of Explainability Technique



**LECTURE
02**

Ch.2 Explaining explainable AI

Ch.2 Explaining explainable AI

2. WHAT IS AN EXPLANATION?

There are nearly as many definitions of explanations as articles about the topic. Philosophers have framed explanation as a form of deductive inference, a form of causal reasoning, or a form of statistical inference.

For instance, a risk manager might ask why a data scientist believes that a model will work well once it is deployed, and a user might ask why they should take the action of accepting an algorithm's recommendation.

In fact, there are three different types of questions that the explaineer might have, which correspond to three different types of reasons.

External reasons ask why an action occurred—independent of the reasons that motivated the agent.

LECTURE
02

Ch.2 Explaining explainable AI

2.1 Explanations of Internal Reasons

Sometimes when an explainee asks for an explanation of an outcome, they may want to understand the beliefs and desires that lead to it.

In the realm of AI, one category of explanation is reasons why the software or artificial agent took some action (such as making a particular prediction) based on its information, beliefs, and goals. Explanations that provide the relative importance of variables—such as using payment and credit usage patterns and ratios like age-to-income for predicting loan defaults—are one example of internal explanations.

They explain an outcome (such as a recommendation that a person's application for credit be declined) based on the AI system's information (such as their income and debt ratios) and goals (such as meeting business goals and respecting the financial institution's risk appetite). Three important categories of techniques to generate internal explanations are feature contributions, examples, and counterfactuals.

This explanation is focused on an aspect of the input to a model rather than its processing logic but is still focused on the internals of a system.

Ch.2 Explaining explainable AI

2.2 Explanations of External Reasons

People sometimes hold mistaken beliefs, which could make internal explanations unsatisfactory. Asking why it failed to properly load the truck is a request for an external explanation that goes beyond the model's internal states. A real-world example of this type of external explanation can be found in the analysis of an early medical AI system for evaluating the risk of patients with pneumonia.

An external explanation of why an AI model recommended an asthmatic with pneumonia be deprioritized for treatment would need to include information outside of the algorithm, such as what data was (or was not) collected and what assumptions were in place.

When a regulatory body asks why the AI was placing and cancelling orders, the explanation might have to come from an analysis that steps outside of the algorithm to consider potential issues with the data, design, or implementation. An important category of external explanation-generation techniques is traceability tools that describe and track how a system was trained, how it was evaluated and tested, and how it is monitored.

LECTURE
02

Ch.2 Explaining explainable AI

Technology Readiness Levels are meant to be a principled process to ensure robust systems, inspired by practices in safety-critical contexts such as spacecraft systems, which could support the creation of external explanations. Explanations that surface relevant examples can also offer valuable external explanations.

For instance, a counterfactual explanation around loan approval could be generated to evaluate a model's reaction to specific low and high values of certain features and these explanations would hint at both internal and external reasons that are pushing the model to this decision.

Ch.2 Explaining explainable AI

2.3 Explanations of Normative Reasons

The term 'normative reason' comes from the notion that there are norms, principles, or codes that make it right or wrong to do certain things.

Business users of AI systems may also be interested in normative explanations.

Managers who fail to ask about whether AI systems have been properly trained to achieve norms such as suitability might be creating legal risks related to negligence, and thus also be interested in normative explanations.

Ch.2 Explaining explainable AI

2.4 Types of Explainability Technique

There are many specific techniques that have been developed to answer these different desires, and any attempt to comprehensively list explainability techniques will likely be outdated soon after it is published.

Researchers have developed explanation-generation techniques that vary depending on the model being explained (such as neural networks, Bayesian networks, expert systems, or model-agnostic approaches), the type of input data and explanation output format (numerical, textual, visual, rule-based, or mixed format), what type of problem the model is being used to solve (such as classification or regression), and whether the technique is applied during or after training.

Some techniques enable developers to construct several simpler models that each approximate parts of the more complex model (such as subsections of the data or input space), while others aim to explain individual predictions rather than attempting to create any kind of a surrogate for the whole or even parts of the complex model.

LECTURE
02

Ch.2 Explaining explainable AI

Another critical distinction in considering different explanation-generation techniques is whether the result will be an AI system that is transparent, interpretable, or comprehensible. For instance, a trading algorithm might use some recognizable features (such as 'liquidity' or 'ratio of short- to long-term traders') to plan an execution strategy, but other features may be mathematical abstractions with no clear human analogue. To be a comprehensible system, investigators must be able to map the relationships between inputs and outputs using concepts (perhaps such as 'credit risk', 'liquidity risk', 'duration risk' or 'economic cycles') that correspond to the investigator's existing ways of structuring knowledge about the world.

This is one example of the larger point that, because the benefits of explainability are mediated by the reactions of the people who receive them, it is important for financial institutions to clearly understand the different personas who might be requesting and receiving explanations.

3. WHY ARE EXPLANATIONS NEEDED?

3.1 Direct Obligations to Consumers

3.2 Justifying the Basis for Conduct

3.3 Rebutting Negligence Claims

3.4 Algorithmic Auditing

3.5 Appropriately Calibrating Trust to Manage Risk

3.6 Sustaining High Performance in Semi-automated Contexts

**LECTURE
02**



Ch.2 Explaining explainable AI

LECTURE
02

Ch.2 Explaining explainable AI

3. WHY ARE EXPLANATIONS NEEDED?

There are at least nine different types of stakeholders within and outside of a financial institution who might want explanations. These are:

1. the subjects of algorithmic decisions (such as ordinary consumers);
2. the data scientists and designers who develop the model, system, and interfaces;
3. business users of the outputs of an algorithm (such as front-office employees);
4. managers and executives who are responsible for the business outcomes and risks tied to use of the model;
5. model checkers within the business (such as other data scientists or an analytics center of excellence) who act as the first line of defense in risk management;
6. independent model checkers in risk and compliance functions who act as the second line of defense for risk management;
7. external auditors who act as the third line of risk management defense;
8. conduct regulators focused on enforcing specific rules (such as fair lending or securities laws);
9. prudential regulators focused on the safety and soundness of the overall financial sector.

LECTURE
02

Ch.2 Explaining explainable AI

These different personas will have different motivations for requesting explanations, and different stakeholders may be looking for different levels or types of insight.

For them, explanations are needed to help calibrate trust and manage risks. Commentaries on explainability often focus on one or a few specific motivations for providing explanations, often the direct obligation to provide explanations to consumers created by some legal regimes.

Financial institutions deploying AI should take a holistic view of the need for explanations, both to ensure these varied risks are well managed, but also to maximize the value that can be created with AI systems.

LECTURE
02

Ch.2 Explaining explainable AI

This implies that all multinational financial institutions, as well as domestically focused financial institutions in the countries cited above, may have explicit obligations to provide explanations based on GDPR. At a high level, the logic of explainability in GDPR is that;

- (a) the regulation prohibits automated processing for significant decisions unless there are safeguards in place;
- (b) one such safeguard is the right of data subjects to receive explanations of how and on what basis the decision was made;
- (c) these explanations are prerequisites for subjects to effectively exercise their other rights (such as contesting a decision); and
- (d) explanations are also prerequisites for systemic (or institutional) controls on automated decision-making, such as algorithmic auditing.

Specifically, GDPR Article 22 states that individuals ‘shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her’ unless there is a suitable basis for processing and there are ‘suitable measures to safeguard the data subject’s rights.’

LECTURE
02

Ch.2 Explaining explainable AI

The Guidelines accompanying GDPR specify that decisions which ‘affect someone’s financial circumstances’—such as approving or denying credit—would likely qualify as significantly affecting someone, and many other financial products that are not explicitly named in the Guidelines could qualify.

GDPR Article 15 also states that if there is solely automated decision-making that falls within Article 22, then the data subjects have the right to access information about ‘the logic involved, as well as the significance and the envisaged consequences of such processing.’ Scholars have also argued that a coherent reading of the two articles implies that if ‘the weights and factors’ used in the decision are necessary to check its accuracy and to potentially challenge its correctness, then such information needs to be part of a valid explanation, not just the general structure of the algorithm.

Thus, GDPR (and related laws) create direct obligations to explain the logic, features, and consequences of AI systems’ decisions.

LECTURE
02

Ch.2 Explaining explainable AI

3.2 Justifying the Basis for Conduct

While GDPR and related laws apply across industries, legislation specific to the financial sector can also create demand for explanation in order to show that a financial institution's conduct that was based on an AI system had an appropriate basis. These might be internal explanations, such as a list that ranks the relative importance of features, or more external explanations, such as the procedures that were used to test and validate the models' assumptions and outputs.

Ch.2 Explaining explainable AI

3.3 Rebutting Negligence Claims

The need for explanations to satisfy conduct basis tests will emerge from industry-specific regulation, such as securities law, but the broader legal category of negligence can also create a need for explanations of AI systems.

Scholars have argued that artificial agents (AI systems) should be treated as legal agents for the purposes of negligent supervision claims, and the general regulatory principle of functional equivalence suggests that ‘a person using a technology which has a certain degree of autonomy should not be less accountable for ensuing harm than if said harm had been caused by a human auxiliary.’ We believe that courts, at least in the US and Europe, may be open to allowing liability for negligent training of an artificial agent because this incentivizes better care in design and programming for the system developer. If this type of recommendation is widely adopted, financial institutions will want to be able to rebut this presumption by having logging and other system-level explanations of behaviour. It should also be noted that when accidents involve semi-automated systems, there is a tendency to shift the focus from systems to people (‘operator error’) even when the system is advertised as being highly resilient.

In using explanations to protect against or assess negligence claims, it is important that financial institutions, regulators, and courts consider the true causes of harm, which will often involve complex interactions between users, models, interaction design, and other factors, rather than focusing explanations solely on the mistakes made by people.

Ch.2 Explaining explainable AI

3.4 Algorithmic Auditing

Claims based on GDPR-style rights, conduct requirements, and negligence are all forms of individual rights. Since 2019, American lawmakers have been considering legislation that would require assessments of high-risk automated-decision systems that would include 'a detailed description of the automated decision system, its design, its training, data, and its purpose.' European lawmakers have also suggested that data controllers explore 'certification mechanisms for processing operations.' While some smaller countries, such as Malta, are developing certification frameworks, there is no consensus even within Europe on what certification of AI systems. Progress has been made on techniques to generate evidence for claims about AI systems, such as generating test inputs to identify corner cases or identifying situations in which incorrect (but potentially plausible) weights could be loaded, which this chapter would categorize as types of internal and external explanations respectively. Therefore, financial institutions should consider how certification and algorithmic audits might both require some kind of explanations to be generated.

Ch.2 Explaining explainable AI

3.5 Appropriately Calibrating Trust to Manage Risk

Apart from legal risks, the use of AI systems may create operational risks that can be managed using explanations. But because the managers could not explain the rationale behind the model's recommendations, they disregarded them.' In this, and similar situations, providing explanations could reduce the risk of disuse. Misuse refers to over-reliance on automation, either when the human operator fails to notice an error or follows an incorrect recommendation.

LECTURE
02

Ch.2 Explaining explainable AI

Organizations can also correct over-reliance bias by using explanations to appropriately calibrate people's trust with AI systems' actual trustworthiness—which in some cases might imply lowering their reliance on the system based on known limitations. Without explicitly mandating the use of explanations to calibrate trust in this way, key regulators, such as the US Federal Reserve and UK Financial Stability Board, have highlighted the links between a lack of interpretability and risk in the financial sector at the macro level and the individual firm level. Although they are rarely framed explicitly in this way, drift detection is a type of external explanation: these techniques focus on identifying divergences between what a model has learned and the actual state of relationships in the world.

Ch.2 Explaining explainable AI

3.6 Sustaining High Performance in Semi-automated Contexts

Some AI applications, such as high-frequency trading, operate at a timescale that makes human involvement in decision-making impractical. However, most applications of AI in financial services are semi-automated in the sense that humans still play an important role either in making decisions (often referred to as 'human in the loop' processing) or supervising algorithms (referred to as 'human over the loop' processing).

Research has found that robo-analysts produce a more balanced distribution of buy, hold, and sell recommendations, and that 'portfolios formed based on the buy recommendations of robo-analysts appear to outperform those of human analysts.' The performance of a joint human-AI team increases when users have explanations, like the artificial agent's reasoning, rationale, or level of uncertainty. Explanations can increase the rate at which users accept the advice given, or action chosen, by an automated system. Perhaps, more importantly, explanations can help signal performance issues that should decrease operators' trust and reliance on the system.

4. HOW TO MAKE EXPLANATIONS EFFECTIVE

- 4.1 Selective**
- 4.2 Robust**
- 4.3 Complete**
- 4.4 Sound**
- 4.5 Actionable**



**LECTURE
02**

Ch.2 Explaining explainable AI

Ch.2 Explaining explainable AI

4. HOW TO MAKE EXPLANATIONS EFFECTIVE

In aiming to make a system more interpretable or more comprehensible, there are several features of explanations that are beneficial. This is based on the underlying assumption that the policymakers, regulators, and risk managers who have created the direct and indirect obligations for explanations want these explanations to be both true and useful.

To be useful, AI explanations should correspond to the implicit 'rules' of how humans explain things to each other, including both people's ordinary social expectations, and their predictable cognitive biases.

LECTURE
02

Ch.2 Explaining explainable AI

4.1 Selective

People prefer to receive explanations that cite fewer causes, so explanations are better if they are tailored to provide a specific and limited amount of information based on each explainee's expectations. Perhaps the consumer wonders why a stock went down rather than up, an analyst might wonder why the volatility was higher than expected, and a regulator might be interested in whether the process was compliant or not. On the other hand, it may be more useful to specifically explain what would need to change ('even if you trebled your salary, you would still not get the loan') rather than a bare cause ('your salary was insufficiently high'), but it is important that these counterfactual changes remain plausible.

LECTURE
02

Ch.2 Explaining explainable AI

4.2 Robust

An explanation is robust to the extent that the causes it identifies remain constant as conditions change. By design, this explanation simplifies a more complex model, but this means that input-output pairs that seem similar to a human observer may have significantly different explanations (i.e., lists of features that matter) depending on the model's higher-dimensional geometry. In a financial context this could mean that an AI model used to value a house might explain that one house's valuation was lowered because it was an older home and local zoning was mostly for large lots, while a very similar house's valuation was increased by these same factors. In contrast, explanations that operate at a higher level of abstraction (such as concepts rather than individual pixels) are more likely to be robust to small changes and comprehensible to explainees who have similar mental concepts.

LECTURE
02

Ch.2 Explaining explainable AI

4.3 Complete

While robustness asks whether an explanation is appropriately agnostic to irrelevant changes, completeness asks whether how well an explanation of a point or portion of input space covers the whole underlying predictive model.

Generating more complete explanations aligns with people's general preferences for receiving explanations based on causes that are necessary, not just sufficient. The amount of content in an explanation can vary significantly and be quite large if, for instance, the explanations are targeted at an expert user inside of an organization rather than an external customer. That is, if a business outcome is the product of two or more models, the explanations of those component models may need to be computed and combined to produce a complete explanation for the overall decision.

Ch.2 Explaining explainable AI

4.4 Sound

In some situations, predictions and explanations are generated by the same algorithm, but in other situations a secondary (surrogate) model may be used to explain a primary model. Soundness is especially important when the explaineer's goal is to identify and fix bugs or errors in the model to ensure they are building an accurate internal representation of what the predictive algorithm is doing.

4.5 Actionable

Designers of AI interfaces have highlighted the importance of making explanations interactive because performance scores (like model accuracy) are rarely enough to produce the desired levels of trust and adoption.

One of the most important potential benefits of an explanation is that it enables the explaineer to intervene in some way, by taking some action after receiving the explanation, such as challenging the decision or requesting human intervention (as envisaged in GDPR and related laws) or improving the model itself.

5. CHALLENGES

5.1 The Challenge of Siloed Expertise

5.2 The Challenge of Limited Guidance from Supervisory Bodies

5.3 The Challenge of Explaining Machine-learned Reasoning

**LECTURE
02**



Ch.2 Explaining explainable AI

Ch.2 Explaining explainable AI

5. CHALLENGES

There are many challenges to producing explanations that are selective, robust, complete, sound, and actionable for the many different types of AI models, data domains, and user personas.

5.1 The Challenge of Siloed Expertise

Since explanations are fundamentally a social undertaking involving human agents, data scientists need to collaborate with designers, psychologists, and other experts outside of math and computer science in order to make better choices in system design. For example, incorporating the explaineer's mental model as part of the explainability system to make it appropriately selective requires a model of human cognition that is at least fit for purpose. Some work has been done at the intersection of these fields such as testing users' preferences for different combinations of simplicity, completeness and soundness.

For instance, one study presented users with two formats of explanations when a model misclassified an input image: one format presented example inputs that had been classified correctly, the other format presented example images of the classes that the model believed were most similar to the input image. For instance, metrics to measure the quality of explanations are still being developed, which leaves developers without an independent, objective way to consider and compare different design choices.

Ch.2 Explaining explainable AI

5.2 The Challenge of Limited Guidance from Supervisory Bodies

The subtle and sometimes highly technical relationships between causality, reasons, and explanations in algorithmic systems can create issues for regulatory bodies that are more familiar with human agents and more deterministic software. However, the presence or absence of variables (or close proxies) in a model that indicates a person's membership in a protected class does not indicate the presence or absence of a causal connection between a decision and class membership, so this proposal might accidentally provide a liability shield to discriminatory algorithms based on a misunderstanding of what could be inferred from their inner logic. The Guidelines on what constitutes 'meaningful information about the logic involved' in processing clarify that the controller does not necessarily need to provide a complex explanation of the algorithms used, but should provide information that is 'sufficiently comprehensive for the data subject to understand the reasons for the decision.' What constitutes 'sufficiently comprehensive' is likely context-specific, which leaves financial institutions with a gap in guidance. Regulators like the Monetary Authority of Singapore are developing detailed guidance and case studies on explainability in financial services. These types of gaps are understandable given the rapid development of AI technology, but the lack of sector-specific guidance on what counts as sufficient explanations is a challenge for financial institutions today that cannot be overcome purely through internal efforts.

Ch.2 Explaining explainable AI

5.3 The Challenge of Explaining Machine-learned Reasoning

Unlike humans, AI systems draw on a limited experience from the training set to establish their logic. This might be predictive in the training set and be a true explanation of the model's reasoning, but nonetheless fail to convince a human explaineer—as it would, correctly, suggest that the model has failed to learn a generalizable classification function.

6. RECOMMENDATIONS TO FINANCIAL INSTITUTIONS ON CREATING EXPLAINABLE AI

6.1 Catalogue Your Explicit and Implicit Explanation Obligations

6.2 Bring Your Developers, Lawyers, Risk Managers and Designers Together for Solutioning

6.3 Develop Competencies for Both Individual- and Institutional- level Explanations

6.4 Embed Explanation Competencies into the Broader Model Risk Management and Governance Framework

6.5 Measure and Manage the Impact of Providing Explanations

**LECTURE
02**



Ch.2 Explaining explainable AI

Ch.2 Explaining explainable AI

6. RECOMMENDATIONS TO FINANCIAL INSTITUTIONS ON CREATING EXPLAINABLE AI

6.1 Catalogue Your Explicit and Implicit Explanation Obligations

For any AI system, there are at least six potential motivations for providing explanations: meeting obligations to individual rights-holders; justifying conduct (such as why a specific product was recommended); rebutting claims that the organization was negligent in training or monitoring the AI system; enabling institutional safeguards such as algorithmic audits; managing risk by helping users and other stakeholders appropriately calibrate their trust in the system; and providing the information needed by human collaborators to sustain high performance in semi-automated settings. Rather than jumping straight into picking among the many available AI techniques, financial institutions should understand the underlying motivations of the related legislation, standard, guidance or internal policy and whether it is best served by an internal, external, or normative explanation. Designers, data scientists, engineers, and product managers (referred to below collectively as 'AI builders') should identify and document explanation requirements early, ideally in the conceptual design phase for new AI systems.

LECTURE
02

Ch.2 Explaining explainable AI

6.2 Bring Your Developers, Lawyers, Risk Managers and Designers Together for Solutioning

Once the various explanation needs are understood, it is important to bring together the many different kinds of expertise required to build safe and effective socio-technical systems.

Finding the right solution requires experts that can ask and answer a diverse set of questions, including human factors considerations (e.g., 'What information will the user be able to absorb and act on?'), operational risks (e.g., 'What data or vulnerability do we risk exposing through explanations?'), legal strategies (e.g., 'Will having or not having certain information help if an adverse event occurs?'), and technical theorizing ('What assumptions is the explanation technique making, and are these true in this case?'). Effective explanations should ideally be integrated early into the system design rather than be added as an extra feature late in development. The need to generate, store, and reproduce explanations and enable users to interact with them should be influenced by the constraints of the production system (such as speed, memory and security), and should in turn impact choices about model and system architecture. The risk can also be reduced by avoiding the temptation to incorporate the very latest advances from academia, and instead using both AI frameworks and explainability algorithms that have been more fully tested in practical applications.

LECTURE
02

Ch.2 Explaining explainable AI

6.3 Develop Competencies for Both Individual- and Institutional-level Explanations

In designing explanations of an AI system, organizations with strong user-centric design practices may be likely to focus on individual explanations for consumers. For instance, model cards and other logging or traceability-style explanations are highly useful for algorithmic audits, and tools that let developers and auditors scientifically probe a model (such as by systematically recording the effects of small changes in inputs or training) are useful for debugging and making AI systems more robust and resilient.

LECTURE
02

Ch.2 Explaining explainable AI

6.4 Embed Explanation Competencies into the Broader Model Risk Management and Governance Framework

AI teams focused on innovation in financial institutions are sometimes separated from business groups and even other analytics or data-science teams that are building more traditional models fully within financial institutions' model governance framework.

While this might appear to accelerate innovation, it risks limiting AI to internal labs or proof-of-concepts if the models that are developed are prohibited or severely delayed by the model risk management process. To avoid this, AI builders should be aware of the type of validation and verification that is required for deployment to production, and the validation and verification teams should be equipped with explainability tools to help with their work. Risks uncovered through explainability work (such as how well the model is likely to perform through the business cycle, or the disparate impact on specific groups due to specific choices of business KPI) should be integrated with risk reporting.

Ch.2 Explaining explainable AI

6.5 Measure and Manage the Impact of Providing Explanations

More comprehensive risk reporting is one positive impact of increasing model explainability, but it is far from the only one. Because building the capacity for explainability requires investment, it is also important to track and quantify the returns these investments are generating beyond compliance.

This is an area where AI practitioners from industry can take the lead in filling a gap in academic research on testing explainability with real-world users and setting. A/B testing the impact of different explainability designs (such as how much or what type of information is provided), measuring developers' usage of explainability-based debugging tools, and instrumenting systems to identify spurious learned relationships are a few examples of how the impact of explanations can be measured. Anecdotally, we believe many financial institutions are under-investing in explainability capabilities compared with the potential benefits, and more measurement of explanations' impact will lead to more investment.

Ch.2 Explaining explainable AI

LECTURE 02

7. Conclusion

This chapter explore how the Financial institutions have an acute need for AI systems that are explainable to ensure they are meeting their obligations to customers, shareholders and regulators. Consumers' right to have individual algorithmic decisions explained is only one source of these obligations; institutional motivations, such as effective legal and operational risk management and maximizing the performance of human-AI teams, are equally important reasons to want explanations. What different stakeholders want from explanations can range from understanding the internal logic of a model, to how a system (including the data generation and consumption processes) is learning patterns that fail to reflect reality, to whether a decision or process can be ethically justified.

Artificial Intelligence for Financial Technology

References:

Required Reference:

Remolina, N. and Gurrea-Martinez, A. eds., 2023. Artificial intelligence in finance: Challenges, opportunities and regulatory developments. Edward Elgar.

Recommended Readings

Chishti, S., 2020. The AI book: the artificial intelligence handbook for investors, entrepreneurs and fintech visionaries. John Wiley & Sons.

The End of Lecture

Thank You

Dr. Abeir Moussa

www.midocean.km

