afme/
Finance for Europe

# Artificial Intelligence and Machine Learning in Capital Markets

## Considerations for a Broad Framework for Transparency

September 2019

Association for Financial Markets in Europe
**www.afme.eu**

# Disclaimer

*Artificial Intelligence and Machine Learning in Capital Markets - Considerations for a Broad Framework for Transparency* (the "Report") is intended for general information only and is not intended to be and should not be relied upon as being legal, financial, investment, tax, regulatory business or other professional advice. AFME doesn't represent or warrant that the Report is accurate, suitable or complete and none of AFME, or its respective employees shall have any liability arising from, or relating to, the use of this Report or its contents.

Your receipt of this document is subject to paragraphs 3, 4, 5, 9, 10, 11 and 13 of the Terms of Use which are applicable to AFME's website (available at http://www.afme.eu/en/about-us/terms-conditions) and, for the purposes of such Terms of Use, this document shall be considered a "Material" (regardless of whether you have received or accessed it via AFME's website or otherwise).

September 2019

# Executive Summary

As the adoption of Artificial Intelligence (AI) and Machine Learning (ML) in capital markets continues at pace, attention is increasingly being focused on how capital markets firms can demonstrate a responsible approach to their use of the technology. This white paper has been developed by AFME's AI Task Force to consider how to approach transparency in AI/ML, which is a key factor in demonstrating and ensuring the safe and effective deployment of trustworthy AI/ML in capital markets. The paper suggests a technology-neutral, principles-based approach to transparency, built around the assumptions used in the development of AI/ML models and testing of those models, to meet stakeholder needs.

In any use of AI/ML, transparency is important to a wide range of stakeholders, as it can demonstrate how an AI/ML model has been developed, how it will be used and monitored, and how it can stand up to scrutiny and challenge. This is crucial for building trust in the technology, both within a firm and with external stakeholders such as clients and regulators.

However, discussions on transparency often quickly develop a specific focus on concepts such as explainability, which involves expressing the complex internal mechanics or workings of an AI/ML model. This is problematic because, while currently available explainability techniques are useful in certain scenarios, in most cases they provide only a partial understanding of complex AI/ML models. In our view therefore, mandating a certain level of accuracy and validity of technical explainability is actually likely to unnecessarily limit the use of the technology, by restricting the breadth and complexity of AI/ML models that can be used and also lead to the provision of 'explanations' that may be misleading and therefore counterproductive.

Instead, we propose that AI/ML transparency should be considered more broadly, as a framework built around (i) qualitative and quantitative assumptions and (ii) testing. Such frameworks should be tailored to the individual risk profile of the AI/ML application and to the needs and knowledge of the various internal and external stakeholders. The framework should also be evaluated and updated throughout the application's lifecycle.

As a heavily regulated sector, capital markets firms recognise that their use of AI/ML must be consistent with their obligations in key areas such as governance, accountability, duty to clients and data protection. The existing regulatory framework for capital markets is largely technology-neutral and principles-based. We suggest that this approach should be maintained for AI/ML, but that a gap analysis should also be performed to ensure that regulations (both existing and new) do not place unnecessary constraints on a firm's use of the technology, or contain granular provisions which may quickly become obsolete as the technology continues to develop.

Our proposed approach to transparency as a way of meeting stakeholder needs fits with our suggested focus on a technology-neutral and principles-based regulatory framework within capital markets. It will enable the demonstration of a responsible and ethical approach to AI/ML and support the development of the technology to the maximum benefit of the industry and its clients. AFME looks forward to working with regulators and the industry to further discuss the issue of transparency in AI/ML and to embed our approach.

# 1. Introduction

AFME established its Task Force on Artificial Intelligence in 2017 with the objectives of increasing awareness of AI/ML in capital markets, supporting the adoption of the technology and contributing to the development of future policy.

This is the third in a series of white papers[1] produced by the Task Force. Our first considered the use-cases, benefits and risks of AI in capital markets, while our second paper explored ethical considerations. This third, more technical, white paper discusses the concept of transparency in AI/ML.

AFME and its members are supportive of the development of AI/ML and are focused on building trust in the technology as its adoption increases across the capital markets industry. This paper considers:

- the importance of transparency for a wide range of stakeholders in the adoption of AI/ML in capital markets;
- why focusing on transparency may be more appropriate than explainability for many AI/ML models;
- how a broad and risk-based transparency framework could meet stakeholder needs and drive trust in the technology; and
- considerations for regulators.

[1] AFME's papers on Artificial Intelligence can be found at https://www.afme.eu/en/divisions-and-committees/technology-and-operations/

# 2. Developing Trusted AI/ML – Transparency and Why it Matters

*AI/ML transparency is important to a wide range of stakeholders, both internal and external to an organisation. It enables them to design, develop, manage, monitor and put their trust in a firm's use of AI/ML.*

Adapting to a Growing Technology

As discussed in our first white paper, 'Artificial Intelligence' and 'Machine Learning' are terms used to reference a broad range of technologies which, although by no means new, are now gaining attention and investment at a significant rate within Europe and globally. Within the broader financial services industry, early results released by the Bank of England[2] in June 2019 of a survey of 200 firms, show that the median firm surveyed is currently deploying six AI/ML applications, with three more expected over the next year and ten over the following three years.

The widespread adoption of any developing technology within a firm will drive consideration of how existing technology governance frameworks can be adapted to ensure that any risks are minimised and mitigated, and that the benefits can be articulated and measured. This is particularly key given the complex and innovative nature of AI/ML: the needs and concerns of a broad range of stakeholders, many of whom may not be practitioners of the technology, must be considered in order to build trust in its outcomes and for its continued deployment.

Many aspects of a firm's existing technology governance framework will be directly transferable to its use of AI/ML, such as approval procedures, or financial accountability. However, AI/ML is a complex and fast-developing technology, about which concerns may arise in relation to control and oversight. Additional care should be taken to ensure that sufficient transparency is provided, in order to address any concerns and provide assurance that it is being used responsibly.

Transparency within AI/ML

For the purposes of this paper, AI/ML transparency can be gained from understanding (i) the assumptions made in the development of AI/ML and (ii) how AI/ML is tested both as part of its initial development and on an ongoing basis. This understanding should be flexible and tailored to the needs of stakeholders and the risks involved in the relevant AI/ML project.

We observe that there is a broad range of related terms for AI/ML, which are defined differently by different writers on the topic. For example, the European Commission High Level Expert Group's 'Guidelines on Trustworthy AI'[3] make reference to 'transparency', 'traceability', 'explainability' and 'interpretability'. Similarly, the 'Declaration on Ethics and Data Protection in Artificial Intelligence'[4] by the International Conference of Data Protection & Privacy Commissioners refers to 'transparency', 'intelligibility' and 'reachability'.

There are no globally agreed definitions for these terms, and at times they are even used interchangeably.

We discuss below why it is important not to restrict our focus to a narrow set of techniques or requirements, which may not be sufficient or appropriate for many applications of AI/ML, and/or may not be able to keep pace with developments in the technology. Therefore we believe that by using the term 'transparency', we are able to propose some considerations for a broad framework which can meet the needs of the various stakeholders involved in any AI/ML project.

Meeting Stakeholders' Needs

As with any technology project, the use of AI/ML within a firm will involve a wide range of stakeholders. Adoption of AI/ML will be dependent on a range of stakeholders' abilities to gain and maintain trust in the firm's ethical and responsible use of the technology, even though their technical understanding of AI/ML may vary. This is where transparency is key: it allows a firm to demonstrate how the AI/ML application has been developed, how it will be used and monitored, and how it can stand up to scrutiny and challenge. Within these broad themes, transparency should meet the varied needs of individual types of stakeholder, both inside and outside the firm. For example, it may:

- contribute to the conception, feasibility analysis and business justification for a AI/ML project;
- assist in the assignment and monitoring of accountability[5];
- provide a suitable basis for the firm's management to sign off and oversee the firm's use of AI/ML;
- enable the firm's technology function to develop, monitor and optimise AI/ML applications;

[2] https://www.bankofengland.co.uk/speech/2019/james-proudman-speech-at-fca-conference-on-governance-in-banking-london
[3] Available at https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[4] Available at https://icdppc.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf
[5] We note, for example, the FCA's recent publication on "Artificial Intelligence in the Boardroom", available at https://www.fca.org.uk/insight/artificial-intelligence-boardroom

- give assurance to internal users of the application of its benefits and performance;
- allow the firm to address concerns that external users or data subjects may have about their interaction with the firm's AI/ML applications;
- demonstrate compliance with ethical and regulatory obligations; and
- enable oversight, auditability and challenge by control functions, e.g. compliance, risk and internal audit.

With these stakeholders and benefits in mind, we will now consider how transparency can be achieved in practice.

# 3. Transparency and Explainability

*Many explainability techniques can be useful for AI/ML, but each has limitations. Any mandate on explainability would ultimately be counterproductive. Rather, explainability techniques should only be used where their methodologies and limitations are understood.*

Within this broad topic of transparency, discussion of how to build trust in AI/ML often focuses on technical explainability and 'explainable' AI (XAI). Explainability is typically referred to as the extent to which the complex internal mechanics or workings of a model can be expressed, while XAI refers to "an AI in which the actions can be easily understood and analysed by humans"[6].

There is no doubt that having the ability to explain the functionalities of any technology application is better than not having this ability. If the option exists to build two models – a first model which is explainable and a second model that is not – then it is generally beneficial to benchmark the models against each other. The second model should only be used where it outperforms the first to a sufficient degree, otherwise it is not bringing enough incremental value to justify the loss of explainability.

However, this approach to explainability is restrictive in that it is often expressed as binary (a model is either explainable or it is not) – it does not allow for the spectrum of AI/ML models that can be used, each of which will have different features and strengths. Equally, it does not allow for developments in explainability techniques and understanding.

Nonetheless, current explainability techniques either can only apply to the simplest AI/ML models, or provide a partial understanding, potentially misleading in its oversimplification, of the AI/ML model. For instance, they reveal very limited information about neural networks[7], a commonly used AI/ML model type and one which has the potential to provide the most benefit when compared to traditional data analytics. This is important, because more complex models generally produce more accurate results, as they are able to use multiple methods of data processing and better uncover non-linear relationships between data points.

Below we consider the challenges related to four examples of conventional AI/ML explainability techniques[8].

## Mechanistic explainability

Explainability in a broader technology context is often understood to mean mechanistic explainability. This may be possible to a limited extent in some AI/ML models for stakeholders with a background in AI/ML. However, the multi-layered and often adaptive nature of AI/ML means that a mechanistic explanation would be too lengthy and complex to be sufficiently comprehensible to a human.

**Mechanistic explainability** is model-specific explanation by design. It expresses the workings of a computer system as: 'input data + process = output'.

Example applications may include linear regressions used in wholesale banking probability of default models or statistical or artisanal (human-reasoned) models used in market risk models such as VaR (value at risk).

Where it is necessary to have an explanation of the mechanism of a technology, it stands that relatively simple, explainable AI/ML models must be used, such as a single, sparse decision tree[9]. Although that may be appropriate in certain use cases, there is generally a trade-off between the simplicity/explainability of an AI/ML model and its performance.

However, we generally believe that placing on a firm's use of AI/ML the requirement that there must be a mechanistic explanation available would significantly limit their ability to develop applications which outperform traditional methods to boost the firm's efficiency and effectiveness.

[6] H. Hagras, "Toward Human-Understandable, Explainable AI", *Computer*, September 2018

[7] "A neural network is a ML system that consists of simple interconnected processing units that are loosely modelled on neurones in the brain" *(for example, an image recognition system that learns to identify a type of image by associating certain features over time).*

[8] For more detail on available techniques, see M. Du, N. Liu and X Hu, "Techniques for Interpretable Machine Learning", Department of Computer Science and Engineering, Texas A&M University, May 2019

[9] A model used to explicitly represent decisions and decision making, Two examples of decision trees used in machine learning are classification and regression trees.

## Feature importance

The key limitation of feature importance is that it does not tell us *how* or *why* a feature is important, i.e. *how*, a model determines its output using features identified as important and *why* we can determine that the 'correct' features have been identified[10].

As with interpretable proxies (below), feature importance is more often implemented at a global level (considering the AI/ML model in its entirety) but can also be used at a local level (focusing on particular outputs of the model). When implemented at a global level, only the average importance of a feature can be determined, which may be of interest but is of limited value when seeking to explain individual outputs. The model can be implemented at a local level, so may illuminate a particular model output. However, the method suffers from other limitations, which it shares with local interpretable proxies and which we discuss below.

**Feature importance** seeks to determine the features, or variables, that have the most impact on the output of the AI/ML application. It works by scrambling/corrupting the data for one feature to see how much of an impact that feature has on the error rate of the model. This can then be repeated for each feature to 'rank' their importance, although this may be time and resource intensive.

Feature importance is particularly useful for rationalising features in the development of an AI/ML model. For example, in developing an AI/ML model which analyses the churn probability of a particular client portfolio, it can be used to identify which characteristics to use in the model. Moreover the important features could be used to reallocate resources or target investment advice.

## Interpretable proxies

In general AI/ML models may benefit from being benchmarked against an interpretable proxy. If the performance of the proxy is near or equal to that of the AI/ML model, the proxy should be used and the AI/ML model discarded, as its complexity brings no additional value. On the other hand, if the proxy's performance is significantly below that of the AI/ML, the proxy is itself of little value as a substitute or an explanation.

Such global proxy models, which seek to replicate the AI/ML model in its entirety, are difficult to create for sophisticated AI/ML models, as these rely on a wide range of variables and interactions between them.

An **interpretable proxy (or surrogate)** involves training an interpretable model to approximate the output of the more complex AI/ML model and then assuming that the two share a common mechanism of action.

For example, a proxy might be created as a sparse (simple) decision tree, or as a linear regression, which attempts to determine the relationship between the variables that an AI/ML model is analysing. Such models are comprehensible to humans.

Local proxy models, which approximate the AI/ML behaviour for a particular instance, have been proposed to understand particular model outputs. These can allow a human to surmise how the model works in a locality, by examining an interpretable model that mimics it well in that locality. However, they suffer from several drawbacks:

1. they are heavily dependent on the point and locality chosen.
2. they are demonstrably fragile, meaning AI/ML models that are similar and behave similarly can generate different local proxies.
3. they must be constrained to be interpretable, so they may produce an 'explanation' that looks questionable because it is an explanation that is inevitably limited by that constraint. For example, the number of variables used in a proxy model would need to be a few less (often far less) than the number of variables in the relevant AI/ML model. It therefore follows that any explanation produced by the (constrained) proxy would be impacted and may appear questionable *because* of this constraint.
4. AI/ML models can be significantly non-linear even in tight localities (i.e. for inputs and outputs very close to the one being 'explained'). This makes it challenging, if not impossible at times, for a linear model (a proxy) to explain accurately how the more complex non-linear model is working, even if the proxy focuses on a limited locality.

[10] This describes techniques such as image analysis, but not predictive analytics or reward optimisation

## Counterfactual explanations

Counterfactual explanations' are generally useful when they identify areas of concern, but where they do not identify an area of concern, such explanations are often of limited value, since the number of possible explanations is high and grows sharply with the model's complexity. Moreover, counterfactual explanations miss complex interactions that can be significant in the model.

So a more direct approach would be to test for the particular counterfactual of concern. To take the example on the right, the explanation might equally have been "all else (including Q1 profits) being equal, if Corporate Y was not still without a CEO, the AI/ML application would not have recommended selling Corporate Y's shares".

A **counterfactual explanation** examines different permutations of the AI/ML input data to determine what changes are required to produce a different outcome. For example, "all else being equal, if their Q1 profits were $X higher, the AI/ML application would not have recommended selling shares in Corporate Y".

Counterfactual explainability can be useful where it identifies a scenario of concern, i.e. a counterfactual scenario that should not have led to a different decision (e.g. an investment recommendation would have been different if, all else being equal, the geographical location or communication channel were different).

The four conventional explainability techniques we describe above are useful in certain scenarios. An example would be a use case where both feature importance and counterfactual explanation techniques are used to explain the outcome of an AI/ML model. Hypothetically, this could involve a targeted search for counterfactual statements, restricted to globally important features. This approach would be used to identify the top globally important features that, if different, could have changed an outcome. All such counterfactual statements will be accurate and important, in the sense that the model considers those features important on average. Also, they resemble how humans explain their decisions.

But as argued above, one should proceed with awareness that (i) globally important features have by definition been averaged, so this approach can miss features that are important off the average; (ii) the number of counterfactual scenarios grows sharply with feature count, so this approach will provide only a (potentially narrow) subset of them; (iii) we would miss correlations — when two or more features, that do not matter as much individually, matter a lot together — especially non-linear correlations, which can be important for more complex models; and (iv) the absence of an offending counterfactual cannot be taken as evidence that it does not exist. In some situations, such counterfactuals may nonetheless suffice, but this judgement must be rendered in full view of the caveats.

In general therefore, these explainability techniques should only be used where their methodologies and limitations are understood by the stakeholders who will receive their 'explanations'.[11] Although it is possible that further research into explainability will significantly enhance such techniques[12], it can therefore be seen that mandating technical explainability today could (a) significantly limit the complexity and accuracy of AI/ML models that can be used, and/or (b) result in apparent 'explanations' that can be misleading. Additionally, such a requirement would be burdensome given than these techniques bring significant cost for their limited benefits and may require additional compliance controls to those for the original AI/ML model.

Furthermore, a reliance on explainability concentrates focus on understanding the inner workings of a particular technology, rather than on requiring humans to demonstrate an appropriate basis for its use. The latter should apply regardless of the type of technology, to ensure that the firm's use of the technology complies with its own policies and standards, and with any regulatory obligations. This is also in line with our belief that AI/ML should generally augment, rather than replace human activity, and with the growing regulatory focus on individual accountability and conduct[13].

Recognising these challenges, we believe that relying on explainability should not be a pre-condition for firms' use of AI/ML within capital markets. Instead, we propose considerations for a wider framework of transparency, tailored to the risks of each application, which can be used to meet stakeholder needs and ensure the responsible deployment of AI/ML for the maximum benefit.

[11] It is problematic to "present a simplified description of a complex system in order to increase trust if the limitations of the simplified description cannot be understood by users… such explanations are inherently misleading" Leilani Gilpin, David Bau, Ben Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning" 2019

[12] AFME and its members are strongly supportive of further research in this area.

[13] For example, there are individual accountability regimes (in force/planned) in jurisdictions such as Australia, Hong Kong, Ireland, Malaysia, Singapore and the UK

# 4. Building Trust Through Transparency

*A risk-based transparency framework built on assumptions and testing should be developed, tailored to the individual needs of the stakeholders involved.*

As outlined in Section 2, stakeholders in an AI/ML project will have different reasons to require an understanding of what the project is designed to do, and whether it is performing well. A reliance on explainability alone cannot meet these needs. We therefore suggest that a wider transparency framework is a more suitable solution.

Such a framework allows suitable oversight and control of the AI/ML model throughout its lifecycle, and can be tailored to ensure that it gives the right level of detail for the different stakeholders and purposes. We suggest that such a framework should be tailored to the stakeholders in any given AI/ML project and then built around two key elements: (i) assumptions and (ii) testing.

## Identification of Stakeholders

As we note above, transparency is key for meeting the needs of the wide range of stakeholders involved in any technology project. Therefore the first step in considering how to approach a transparency framework should be the identification of the various categories of stakeholder and what their interest in the project will be[14]. The suggested list of stakeholder categories in Section 2 should be considered non-exhaustive, as each AI/ML project will have its own requirements.

Once the stakeholders and their needs have been identified, firms should consider how those needs can be met using assumptions and testing as set out below. It is likely that the different categories of stakeholder will need differing levels of detail about the approach taken by the firm in relation to each. For example, some may require technical or coding detail, but others may need information in natural language, and some may be more interested in the process and controls, while others focus on understanding an application's outcomes.

## Assumptions

'Assumptions' refer to those elements of the AI/ML development process whose validity is often accepted without further and detailed proof.

There are a number of ways in which both quantitative and qualitative assumptions can manifest throughout the lifecycle of an AI/ML application. These should all be clearly articulated at the outset of any project to create an AI/ML application and then reviewed as appropriate throughout the application's lifecycle.

The below table sets out examples of assumptions, as well as questions to consider in ensuring that the assumptions do not have unintended adverse effects.

| Type of Assumption | Questions to Consider to Mitigate Possible Adverse Effects |
|---|---|
| Methods: assumptions implicit in developer's decisions, e.g. the AI/ML technique chosen | How have the decisions been justified in developing this application? |
| Data: the sufficiency of the data set and its structure (data fields) in the context of the application | How have the data been selected and screened?<br>How has bias within the data been identified and mitigated?<br>Which population does the data set represent? |
| Goals: the alignment of selected optimisation goals with the objectives of a specific application | How have the objectives, e.g. the reward function, of the application been defined?<br>What measures will the application use to determine whether it is meeting its objectives?<br>How will the continuous alignment with those goals be monitored? |
| Results: the adequacy of success criteria by which the application's performance is judged | How will the firm assess the application's output?<br>What are the acceptable tolerances for its performance?<br>How will the output be used and are limits required (e.g. the setting of a cap/floor)? |

[14] The Bank of England's Staff Working Paper No. 816, 'Machine learning explainability in finance: an application to default risk analysis' shows one such analysis. Table 1 of the paper lists some of the types of explanation that may be required by developers, management, conduct/prudential regulators etc.

## Testing

An AI/ML application must also be subject to rigorous testing, both during development and when the application is live. Much of this will be part of firms' existing software testing programmes, and will include testing against different market or system conditions, or to determine the interaction between the application and other systems. The application should also be tested against contrived inputs intended to probe boundaries or uncover unethical behaviour. For example, testing a market risk attribution model should include use of simulations of market turbulence that has occurred in the past, but also future hypothetical scenarios.

## Taking a risk-based approach

As noted in our first white paper, AI/ML can potentially be used in a whole range of functions across capital markets, both to augment existing activities and to perform complex and intensive tasks which would otherwise be impossible to execute. Each use-case will have its own risk profile and key stakeholders, which will need to be mapped out at the start of the project and monitored throughout its lifecycle. The considerations for a transparency framework that we outline above are intended to be adaptable to suit the needs of each use case: i.e. a higher risk model would be set a higher bar on transparency around assumptions and testing.

For example, those who develop AI/ML for algorithmic trading will need to consider the potential risks to clients and markets, ensuring that there is sufficient transparency as to the controls that have been put in place and the testing that is undertaken on the application's performance under stressed market conditions. On the other hand, an AI/ML application designed to manage a non-critical operational process may naturally have a lower risk profile and may call for a different level of transparency around assumptions and testing. There may also be scenarios where it is necessary to limit transparency: for example, when developing AI/ML for the detection of financial crime, a high degree of transparency to a wide range of stakeholders increases the risk that the system could be manipulated or circumvented by those it is seeking to protect against.

# 5. Considerations for Regulators

*AFME supports a technology-neutral and principles-based approach to regulation, in line with our considerations for a transparency framework, rather than prescriptive explainability requirements. A review of the existing capital markets regulatory framework would identify gaps or areas where amendment is needed to support the development of AI/ML.*

The use of AI/ML in capital markets is subject to a number of existing regulatory requirements in areas such as governance, accountability, duty to clients and data protection[15]. These regulatory requirements are largely technology-neutral, applying equally to manual processes as to sophisticated AI/ML systems. Many of these requirements already drive the way that firms are developing and adopting AI, closely linked to the needs of the different stakeholders outlined above.

Given the highly-regulated nature of capital markets, AFME and its members do not believe that it is necessary for regulators to design a new regulatory framework for the use of AI/ML. Instead, we suggest that a gap analysis of existing regulations should be performed, in order to ensure that they are focused on the appropriate outcomes and that it is not unintentionally placing constraints on firms' use and upscaling of AL/ML applications. This may require exploration of how existing rules can be adapted to support the development of the technology to its full potential, both in Europe and globally[16].

In relation specifically to the subjects of explainability and transparency in AI/ML, we suggest that a principles-based, technology-neutral approach should continue to be followed. Maintaining this approach will ensure that firms can be held to high standards without granular rules that quickly become obsolete or impractical as the technology continues to develop. This will also allow senior management within a firm to design policies and procedures tailored to their own businesses and risks, which meet their requirements as the accountable executives for the firm's activities.[17]

The regulatory approach should not set accuracy and validity levels for AI/ML explainability. As we have suggested, this is challenging in all but the simplest uses of AI/ML and could constrain the use and potential benefits of AI/ML. Instead, a broader approach to transparency should focus on the assumptions and testing which allow humans to decide how and when to develop and use AI/ML, and to evidence how decisions contribute to the responsible and ethical deployment of AI/ML in capital markets.

[15] For example, in Europe, the Markets in Financial Instruments Directive and Regulation, the General Data Protection Regulation, or the Capital Requirements Directive.

[16] In this, we look forward to the publication of the recommendations of the European Commission's High Level Expert Group on Regulatory Obstacles in Financial Innovation, which may address issues relevant to AI/ML

[17] We note for example, the principles-based approach suggested by DeNederlandscheBank in its July 2019 paper 'General Principles for the use of Artificial Intelligence in the Financial Sector', where it is stated that "decisions that favour accuracy over traceability and explainability should be well-motivated, documented, and approved as the appropriate level"

# Conclusion

As AI/ML deployment continues at pace within capital markets, it is natural that there should be increasing attention on how firms are ensuring that an appropriate level of governance and oversight is in place, which identifies where existing technology policies and procedures need to be adapted to the unique features of AI/ML.

When considering this, there is often a specific focus on explainability, with the suggestion that an AI/ML model is either explainable or it is not explainable at all. We believe that such a binary approach is not appropriate for categorising AI/ML, as it does not allow for developments in either AI/ML models or explainability techniques. However, we also recognise that conventional explainability techniques may not be particularly useful for many AI/ML applications, providing either very complex mechanistic explanations valid only in a technical context, or partial, even misleading, 'explanations'.

Therefore, we have proposed considerations for a framework built around the broader concept of transparency. This involves identification of the various stakeholders in an AI/ML project and their needs, which should then be met through a structure of (i) qualitative and quantitative assumptions and (ii) testing. Both should be articulated at the start of any AI/ML project, then monitored and adjusted as necessary throughout its lifecycle. This approach can be tailored to the risk profile of each individual AI/ML application, rather than applying 'one size fits all' standards.

AFME's members are committed to developing and deploying AI/ML in a manner that is consistent with their regulatory and ethical obligations. As a highly-regulated industry, capital markets firms are already subject to a broad range of technology-neutral requirements that are directly applicable to their use of AI/ML. We encourage regulators and policymakers to continue this approach, regulating outcomes rather than technologies, which will ensure that regulation is able to keep pace with new technological developments and not place unnecessary obstacles on the industry's use of the technology.

We believe that an AI/ML transparency framework is achievable within the existing rules, laws and regulations applicable to the capital markets industry and will support firms in meeting their regulatory and ethical obligations, and in deploying AI/ML to the maximum benefit for themselves and for clients. We look forward to working with the industry to achieve this aim.

# Annex 1: Glossary of Terms

| Glossary of technical terms used in this paper | |
|---|---|
| Assumption | Assumptions refer to those elements of the AI/ML development process whose validity is often accepted without further and detailed proof. |
| Counterfactual Explanation | A counterfactual explanation suggests how differences in the AI/ML input data or process might produce a different result. |
| Decision Tree | A model used to explicitly represent decisions and decision making, Two examples of decision trees used in machine learning are classification and regression trees. |
| Explainability | Explainability typically refers to the extent to which workings of a model can be understood. |
| Explainable Artificial Intelligence (XAI) | An AI in which the actions can be easily understood and analysed by humans. |
| Feature Importance | Feature importance seeks to determine the features, or variables, that have the most impact on the output of the AI/ML application. |
| Interpretable proxy | An interpretable proxy is a simple and human-comprehensible model which mimics the AI/ML model and thus can be used to infer explanations. |
| Mechanistic Explanation | Mechanistic explainability is model-specific explanation by design. It expresses the workings of a computer system as 'input data + process = output'. |
| Neural Networks | "A neural network is a ML system that consists of simple interconnected processing units that are loosely modelled on neurones in the brain" *(for example, an image recognition system that learns to identify a type of image by associating certain features over time).* |
| Transparency | A clear and risk-based understanding of (i) the assumptions made in the development of AI/ML and (ii) how AI/ML is tested both as part of its initial development and on an ongoing basis. |

# Notes

## Bibliography


European Commission High Level Expert Group, 2019, Guidelines on Trustworthy AI

International Conference of Data Protection & Privacy Commissioners, 2018, Declaration on Ethics and Data Protection in Artificial Intelligence

Organisation for Economic Cooperation and Development (OECD), 2019, Artificial Intelligence in Society

DeNederlandscheBank, 2019, General principles for the use of Artificial Intelligence in the financial sector

P. Bracke, A. Dutta, C. Jung and S. Sen, Bank of England, 2019, Staff Working Paper No. 816 - Machine learning explainability in finance: an application to default risk analysis

M. Du, N. Liu and X Hu, 2019, Techniques for Interpretable Machine Learning

L. Gilpin, D. Bau, B. Yuan, A. Bajwa, M. Specter and L. Kagal, 2018, Explaining Explanations: An Overview of Interpretability of Machine Learning

H. Hagras, 2018, Toward Human-Understandable, Explainable AI


## Contributors to this Paper


We are grateful to our member firms and the individuals who contributed their time and thoughts in producing this paper.


## AFME Technology and Operations

AFME's Technology and Operations Division brings together senior technology and operations leaders to influence and respond to current pan-European market drivers and policy.

The Considerations on the Ethical Use of Artificial Intelligence in capital markets white paper was led by the AFME AI Task Force as an initiative within the broader Technology and Operations Division.

## AFME Contacts


Andrew Harvey
andrew.harvey@afme.eu
+44 (0)20 3828 2694

David Ostojitsch
david.ostojitsch@afme.eu
+44 (0)20 3828 2761

Fiona Willis
fiona.willis@afme.eu
+44 (0)20 3828 2739

# / About AFME

The Association for Financial Markets in Europe (AFME) is the voice of all Europe's wholesale financial markets, providing expertise across a broad range of regulatory and capital markets issues.

We represent the leading global and European banks and other significant capital market players.

We advocate for deep and integrated European capital markets which serve the needs of companies and investors, supporting economic growth and benefiting society.

We aim to act as a bridge between market participants and policy makers across Europe, drawing on our strong and long-standing relationships, our technical knowledge and fact-based work.

## Focus
on a wide range of market, business and prudential issues

## Expertise
deep policy and technical skills

## Strong relationships
with European and global policy makers

## Breadth
broad global and European membership

## Pan-European
organisation and perspective

## Global reach
via the Global Financial Markets Association (GFMA)

afme/
Finance for Europe


**London Office**
39th Floor
25 Canada Square
London, E14 5LQ
United Kingdom
+44 (0)20 3828 2700

**Brussels Office**
Rue de la Loi, 82
1040 Brussels
Belgium
+32 (0)2 788 3971

**Frankfurt Office**
Skyper Villa
Taunusanlage 1
60329 Frankfurt am Main
Germany
+49 (0)69 5050 60590

**Press enquiries**
Rebecca Hansford
Head of Media Relations
rebecca.hansford@afme.eu
+44 (0)20 3828 2693

**Membership**
Elena Travaglini
Head of Membership
elena.travaglini@afme.eu
+44 (0)20 3828 2733

**Follow AFME on Twitter**
@AFME_EU