Perspectives on Explainable AI in The Financial Sector

An exploratory study between banks and supervisory authorities

DeNederlandscheBank
EUROSYSTEEM

iForum discussion | learning | pilots


Introduction


Framework


Use cases


Conclusions
and way forward

Practical information

In 2019 De Nederlandsche Bank (DNB) published the discussion paper "General Principles for the use of Artificial Intelligence in the Financial Sector". One of the central principles of this discussion paper underscores the importance of explainability of AI-driven decisions and model outcomes.

In this publication, we further explore the issue of explainability, and its relevance with regard to the responsible use of AI in the financial sector, based on a number of semi-structured interviews. This exploratory study is the result of a collaboration between DNB, Authority of Financial Markets (AFM), University of Applied Sciences Utrecht (HU), The Dutch Banking Association (NVB), and representatives from three major banks in The Netherlands. A collaboration that was made possible by DNB's iForum.

DNB's iForum aims to create more room for technological innovation within the financial system and does so by developing joint experiments in the areas where technology and supervision meet. With this study, we have therefore outlined the perspectives from the banks as well as the two involved supervisory authorities, on where explainable AI meets supervision.

This exploratory study does not propose or formulate new supervisory policies. Instead, it presents a stock-take of current perspectives, observations regarding potential challenges in the implementation of explainable AI, as well as possible areas for future exploration.

We believe that this study poses several findings that are suitable for further discussion, which will be taken up by the iForum in 2021. Questions regarding this publication can be sent to iForum@dnb.nl.

Management summary

Explainable AI is defined as "A set of capabilities that produces an explanation, in the form of details, reasons, or underlying causes, to make the functioning and/or results of an AI solution sufficiently clear so that it is understandable and addresses stakeholders' concerns".

Explainable AI (xAI) is becoming of utmost importance to ensure a responsible use of artificial intelligence (AI) solutions, which are increasingly applied in financial services and products. Although many reports on the responsible use of AI have stressed the need for xAI, a lack of publications that provide practical guidance or frameworks, to facilitate financial organizations in applying xAI, was found. Therefore, this exploratory study aims to bring together the perspectives from the participating banks as well as two financial supervisory authorities (AFM and DNB) on xAI for AI applications in the field of consumer credit, anti-money laundering (AML), and credit- risk management.

The objective of this publication is threefold: (i) to gain insight and understanding in the most relevant aspects of xAI in finance, (ii) to build common ground, as well as a shared vocabulary, to facilitate future studies and discussions on explainability, and (iii) to identify potential issues that merit further exploration. This is done with the use of a practical framework on xAI, applied to several use cases in the fields mentioned above.

Based on semi- structured interviews with representatives of the participating banks and two supervisory authorities regarding the application of the xAI framework on different use cases (in the field of consumer credit, anti-money laundering, and credit-risk management) the main findings from this exploratory study are as follows:

- There appears to be a disparity between views of the supervisory authorities and the participating banks, regarding the desired scope of explainability required for AI solutions in banking.
- There is a need for enhanced cooperation between supervisory authorities, and alignment of supervisory authorities on a national and EU level on the topic of AI and explainability of AI models.
- It would be beneficial to create a space of common trust in order to discuss and share (ideas for) AI-driven innovations.
- Relevant aspects of explainability were found to be highly contextual and to vary per AI use case. The framework as used in this study was regarded in the interviews to be a potentially valuable starting point, but not as a comprehensive approach, to consider suitable xAI elements for the considered use cases.

Suggestions for further exploration of these findings can be found in chapter 4 ("Conclusions and way forward") of this publication.

1 Introduction

In recent years increasingly powerful – but often also increasingly complex – artificial intelligence (AI) models have become available.[1] The use of these models in a growing number of practical applications has sparked a discussion on the need for explainability. On 21 April this year, the European Commission published their proposal for a European AI regulation, which also features requirements regarding explainability in certain high-risk applications such as consumer credit scoring.[2]

This discussion is driven by the notion that many such AI applications have, or potentially have, a material impact on the wellbeing of individuals, groups, and on society at large. Consequently, there is a need to understand how certain AI applications actually work, and how they arrive at outcomes and decisions based on these outcomes. There are numerous examples of situations in the financial sector where comprehensive understanding seems to be warranted, such as in the case of a failed loan application, the assessment of a new internal ratings-based model, or an unusual transaction alert. At the same time, expectations regarding the performance of AI applications are likewise high, and striking the right balance between performance and explainability can present a difficult dilemma.

Explainable AI (or 'xAI') aims to ensure AI can be better understood, by making algorithms and their applications more transparent and less of a "black box".

An improved understanding of the working of these algorithms, helps us to verify them, improve them, and learn from them. Most developments in xAI focus on technical tools for model developers[3] or broader approaches to explainability.[4] Given the paramount importance of explainability, it is no surprise that many reports on the responsible use of AI have stressed the need for xAI. Unfortunately, there seems to be a lack of publications that provide practical guidance or frameworks to facilitate organizations in applying xAI in practice.[5] Furthermore, there does not yet seem to exist a shared vocabulary –in the literature or in the sector– regarding the definitions of and relations between the related (but different) concepts of explainability, interpretability, transparency and traceability.

Finally, a clear taxonomy of the various types of explanations, and how these relate to different potential stakeholders, is not yet readily available. This lack of a solid and practical framework for the financial sector poses a challenge for financial institutions to understand their obligations (regulatory and otherwise) regarding xAI, and how to operationalize them. At the same time, the lack of such a framework also makes it difficult for supervisory authorities to interpret the implications of current regulations regarding transparency and the provision of information, and to communicate their expectations regarding xAI to regulated entities.

1 In this paper we do not propose a strict definition of artificial intelligence, but in the context of this paper generally use this term to refer to systems that use typical artificial intelligence methods (such as deep learning, random forests, and kNN) for machine-learning enabled tasks such as regression, classification, clustering and density estimation.
2 Article 13 of the proposed regulation states that ""High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately".
3 e.g. SHAP, Lundberg & Lee (2017)
4 Miller (2019)
5 Two notable exceptions are the report from the UK Information Commissioner's Office / Alan Turing Institute ("Explaining Decisions made with AI"; May 2020) and the France ACPR discussion paper ("Governance of Artificial Intelligence in Finance"; June 2020)

1.1 Objective

Based on the gaps identified in definitions, concepts and taxonomies for explainability, this exploratory study set out to find common ground (in vocabulary, values, and the practicability of xAI) between supervisory authorities and regulated entities in the financial sector. The point of departure for the study was to exchange ideas and further develop a shared conceptual framework on xAI by applying it to three practical use cases. The findings and observations drawn from the discourses between the participants yielded insight into possible areas for further research and collaboration, and provide guidance for both supervisors and financial institutions in terms of applying xAI in practice. Moreover, the aim was to facilitate the discussion between supervisory authorities (in this case DNB and the Dutch Authority for the Financial Markets) and financial institutions (in this case three banking institutions) regarding the application of AI models, mutual expectations, and the types and degrees of desired explainability for a responsible use of AI in the financial sector.

1.2 What is xAI?

Explainable AI (xAI), also referred to as interpretable or understandable AI, aims to solve the "black box" problem in AI. A typical AI solution utilizes data (e.g. information on a person's financial situation) and produces an outcome (e.g. a risk of default indication). However, in such a process it is not always evident from the output how or why a certain outcome is reached based on the data. Especially when using more complex AI models (e.g. deep neural networks) the process from input to output is practically impossible to interpret in terms understandable by humans, even with knowledge of the inner workings, weightings, and biases of the system. The term xAI encompasses a wide range of solutions that explain why or how an AI model arrives at outcomes or decisions, in general and/or in specific cases.

During this study the following definition of xAI is used: "Given a stakeholder, xAI is a set of capabilities that produces an explanation (in the form of details, reasons, or underlying causes) to make the functioning and/or results of an AI solution sufficiently clear so that it is understandable to that stakeholder and addresses the stakeholder's concerns."

This definition and the conceptual framework used in this preliminary study originate from the paper by van den Berg and Kuiper (2020), which relates the various stakeholders to the types of explanation they require for specific use cases.

1.3 Scope

The shared conceptual framework[6] was applied to three use cases in finance:

1. Consumer credit
2. Anti-money laundering (AML)
3. Credit-risk management

These three use cases each have different domains and stakeholders, and can potentially lead to different interpretations regarding the need for xAI.

1.3.1 Consumer credit

Consumer credit is credit, other than a mortgage loan, which can be used to purchase goods and services. Financial institutions that provide consumer credit have the right and obligation to ensure that the borrower has the capacity to repay the loan. Credit risk assessment processes have evolved from assessments and decisions made by loan officers to more automated decision-making requiring minimal human intervention because of the use of AI.

6 Van den Berg & Kuiper (2020)

1.3.2 Anti-money laundering (AML)

The use case on anti-money laundering has a focus on AI applications which are used to conduct suspicious activity monitoring and transaction monitoring. In this study, we considered a use case with a machine learning aspect used for transaction monitoring, and an AI-solution developed to detect fraudulent activity in corresponding banking transactions.

1.3.3 Credit- risk management

The use case on credit-risk management focusses on internal risk and/or capital requirement models, early warning systems and probability of defaults models, where AI applications are used to improve or replace the currently-used models. The AI-application analyzed is an AIRB (advanced internal rating-based) model for a bank's residential mortgage portfolio - a capital model.

1.4 Structure

This exploratory study has been structured in the following manner:

- First, exploratory discussions were held with the participants on xAI requirements and expectations;
- Second, the supervisory authorities exchanged views with the financial institutions to identify relevant use cases, and to prepare a research methodology (e.g. interview template);
- Third, a series of semi-structured interviews with stakeholders and banks were conducted;
- Fourth, all the participants were provided with written feedback, followed up by a plenary discussion to collectively arrive at main findings.

The semi-structured interviews were the main source of information for the use cases covered in this document. As a starting point, questions were drafted for a range of topics on xAI, but discussion was permitted to develop naturally in the direction deemed most suitable by the persons involved. One interview per use case and institution was conducted (bank or supervisory authority); in some cases, an interview took place to address the issue of xAI in general, in addition to the interview on the specific use cases. For the interview, the use case experts from the institution were invited. These experts either had a technical expertise (i.e. those directly involved with the development) or a more supervising/governing role.

Based on the framework and corresponding whitepaper, a template was drafted which addressed the applicability of the framework to a specific use case. This document was used by the interviewees as a starting point to help them reflect on the framework for each use case.

2 Framework

2.1 Main concepts relevant for xAI

To better understand xAI and its objectives and aims, a clear picture of various associated concepts is essential. A critique on the xAI field expressed by various authors is that xAI is often not clearly defined and discussed without proper understanding of the surrounding concepts and the parties involved.[7] Furthermore, as there are types of model that are inherently more transparent (and less "black box") there is a discussion about whether opaque ML models should be avoided altogether, which can be seen as a different approach to xAI.[8] The most important concepts surrounding xAI are covered below; for a more exhaustive list see the whitepaper.[9]

- **Transparency** is one of the central concepts of xAI. Importantly, the term is used in two distinguishable contexts or manners. 1) Model transparency is the property of a model to be understood by a human as it is, in terms of its general working or design. The opposite of "black-boxness" is (model) transparency.[10] This type of transparency is generally what model developers refer to. 2) Transparency of the implementation of an AI model or system relates to openness and not concealing information for stakeholders. This form of transparency is generally what the colloquial meaning of transparency refers to. This has also been called "process transparency".[11]

- **Interpretability** is closely related to transparency, as in most definitions an interpretable model implies model transparency. Occasionally, interpretability is defined as the capacity of a system to be made understandable via a proper explanation, even if it is not interpretable as it is.

- **Understandability** entails being comprehensible and able to be understood by a specific receiver of the explanation, i.e. a stakeholder. This concept is less ambiguous than e.g. transparency, but underlines the importance of a stakeholder.

- **Stakeholder** refers to the individual or party requiring information often in the form of an explanation.

- **Explainability** entails that an explanation can be formulated. An explanation is contextual, relevant and has the goal of addressing a stakeholder's concern or interest.

- **xAI** as the definition is used for this project: Given a stakeholder, xAI is a set of capabilities that produces an explanation (in the form of details, reasons, or underlying causes) to make the functioning and/or results of an AI solution sufficiently clear so that it is understandable to that stakeholder and addresses the stakeholder's concerns.

7 Lipton (2018)
8 Rudin (2019)
9 Van den Berg & Kuiper (2020)
10 Arrieta et al. (2019)
11 Gerlings et al. (2020)

2.2 Stakeholders in xAI

The following list of stakeholders for AI in the financial sector is proposed:[12]

- End user (external e.g. a customer, or internal)
- Explainer to the end user (e.g. financial advisor or loan officer in the case of a customer)
- AI developer
- Domain expert
- Executive management
- Management responsible for the AI solution (1st line)
- Operational control (2nd line)
- Internal audit (3rd line)
- Regulatory stakeholders and others (e.g. supervisory authorities and auditors)

Stakeholders have various interests, and as such, an explanation can have various goals. For instance, an explanation can have the goal of societal acceptance, regulatory compliance, system development, or aiding in model development. Examples of intermediate goals are model accuracy, interpretability, fidelity, fairness, privacy, usability, reliability/robustness, or scalability.[13] Regardless of the goal, a stakeholder has a selection of specific concerns that can be met by various types of explanation, i.e. that require specific information.

2.3 xAI Framework

In the framework below, relevant stakeholders for a generic AI use case in finance are cross-referenced with the types of explanation potentially required by that stakeholder; i.e. the content of explanation that might address the main concerns of that stakeholder.

The types of explanation cover both local and global explanations in terms of model output and performance, but also extend to processes surrounding the AI such as data selection, design, and governance. As such, it covers concepts such as transparency and explainability in the wider sense generally found in the literature, rather than the more narrow sense only pertaining to the AI model and its direct developers.

12 Depending on the context or use case, the amount and nature of stakeholders may differ
13 See e.g. Arietta et al. (2019)

Figure 3.1 Conceptual xAI framework taken from Van der Berg & Kuiper 2020. The cells with a "■" indicate that this type of content of explanation might be relevant for a specific type of stakeholder

	Type of stakeholder											
Type of explanation	External		Service provider							Other stakeholder (context)		
	End user	End user's external advisor	End user's external advisor	AI developer	Domain expert	Executive mgmnt	Operational mgmt. (1st line)	Operational control (2nd line)	Audit (3rd line)	Regulator A	Regulator B	Regulator C
The reasons, details or underlying causes of a particular outcome	■	■	■									
The data and features used as input to determine a particular outcome	■	■	■	■								
The data used to train and test the AI system				■	■	■	■	■	■			
The performance and accuracy of the AI solution				■	■	■						
The principles, rules, and guidelines used to design and develop the AI solution				■	■	■	■	■				
The process that was used to design, develop, and test the AI solution					■		■	■	■			
The process of how feedback is processed					■		■	■	■			
The process of how explainers are trained					■		■					
The persons involved in design, development and implementation of AI solution					■	■	■	■	■			
The persons accountable for development and use of the AI solution						■	■	■	■			

3 Use cases

This chapter discusses the three use cases from the perspective of the supervisory authorities and the banks:

- Consumer credit
- Credit-risk management
- Anti-money laundering

3.1 Consumer credit

Interviews were conducted with two banks as well as with the AFM for the consumer credit use case.

The first bank provided a case for mortgage lending. The purpose of this AI application is to flag mortgages with traffic-light colors to support middle office employees. The AI application runs in parallel to other, more traditional applications in the mortgage approval and monitoring process. The AI application is rather traditional, based on logistic regression, and uses 10 variables. It differs from a business rules application in that it uses historical data. The bank considered whether the users of the AI application, mid office employees, should be provided with detailed insight into the workings of the AI application. In this particular case, after weighing the degree of information to be provided to these employees, it was decided not to give full insight in order to prevent potential gaming of the system.

The second bank recently introduced an add-on to the traditional loan approval for consumer credit process, with transactional data at the basis. This new AI application runs alongside the traditional application, leading to an ensemble model. The traditional application uses basic data, such as the data a client provides through the application process or data from credit bureaus. The new application is trained and continuously fed with transactional data. The combination of both models resulted in fewer defaults on loans.

For this use-case, model developers are considered the most important stakeholders with regard to explainability. It was observed that it would be possible from a technological point of view to explain the model to customers, although this requires a thorough understanding of which type of narratives would be comprehensible by different consumer groups. This might require an interactive process, which can present a challenging IT problem.

From the perspective of the AFM, the so-called lending standards are the basis for loan approval (or rejection). The AFM assesses lenders for compliance with lending standards. Lenders must comply with the lending standards in 100% of the cases. Therefore, when applying AI, the AI application must at all times be 100% accurate. When a bank uses AI in the context of consumer lending, from a model supervision perspective the AFM is interested in all types of explanations as is shown in figure 4.1, either to be able to assess compliance with the lending standards, or to assess and monitor the whole lending process.

It is observed that there are differences in interpretation regarding the exact scope of explainable AI. From the adopted definition of explainable AI, it is not sufficiently clear whether specific informational needs from the supervisor should be regarded as proper explainability concerns or as instances of regular supervisory information requirements that are not specific for AI.

Figure 4.1 Types of explanation relevant for the supervisor according to the supervisor and the participating banks

Type of explanation	Relevant for supervisor (AFM)	
	According to AFM	According to participating banks
The reasons, details or underlying causes of a particular outcome	■	■ / -
The data and features used as input to determine a particular outcome	■	■
The data used to train and test the AI system	■	■
The performance and accuracy of the AI solution	■	■ / -
The principles, rules, and guidelines used to design and develop the AI solution	■	■
The process that was used to design, develop, and test the AI solution	■	■
The process of how feedback is processed	■	■ / -
The process of how explainers are trained	■	■ / -
The persons involved in design, development and implementation of AI solution	■	■ / -
The persons accountable for development and use of the AI solution	■	■

In summary, for consumer credit, explainability towards consumers (loan applicants) is technologically possible for the evaluated use-cases, but potentially challenging. Furthermore, in most cases there is a human-in-the-loop (the advisor) who provides the customer with information. The lending standards are the basis (and thereby the explanation) for rejection of most loans. However, for cases where an AI application is used to reject cases exceeding the restrictions of lending standards, explanations to consumers might become an issue in the future.

3.2 Credit risk

One bank as well as DNB were interviewed with regard to credit risk. The AI-application of the bank is an AIRB (advanced internal rating-based) model for the bank's residential mortgage portfolio - a capital model. It predicts a probability of default for each mortgage customer and a prediction of loss-given-default for each customer. In essence, the model is simple. It contains only 10-15 variables and is based on logistic regression. There is no interaction with the consumer based on the model. The model is used to calculate capital. The main stakeholders for explanations are the internal first line and the supervisory authority (JST in this case). More advanced AI could probably lead to better performance, but the bank is reluctant to start using these advancements, due to the expected long and time-consuming process to get approval, both internally and externally (JST).

From the interview with DNB, it became apparent that regulations heavily determine the boundaries for what kind of AI-applications can be used. Until now, only logistic regression models are used across all financial institutions participating in the study. Models that are more complex may not meet requirements like traceability and replicability. Another requirement for credit risk models is to demonstrate "experience" in applying a model. In practice, this means that the model must be used as a shadow model for at least three years before approval can be given.

Banks are conducting plentiful research and pilots into AI in credit risk, but the regulations are a limiting factor for further implementation: AI in credit risk currently does not appear to lead to sufficient benefit compared to the challenge of getting its use approved within the current regulatory framework to make it worthwhile. The bank that is the first to implement new AI methods must assume that it takes at least 1.5 years before approval is granted. Figure 4.2 shows the types of explanations relevant for the supervisor, through the eyes of DNB and the participating banks.

Figure 4.2 Types of explanation relevant for the supervisor according to the supervisor and the participating banks

Type of explanation	Relevant for supervisor (JST)	
	According to DNB	According to participating banks
The reasons, details or underlying causes of a particular outcome	■	
The data and features used as input to determine a particular outcome	■	■
The data used to train and test the AI system	■	■
The performance and accuracy of the AI solution	■	
The principles, rules, and guidelines used to design and develop the AI solution	■	■
The process that was used to design, develop, and test the AI solution	■	■
The process of how feedback is processed	■	
The process of how explainers are trained	■	■
The persons involved in design, development and implementation of AI solution	■	
The persons accountable for development and use of the AI solution	■	■
Other: data lineage	■	

To sum up, in credit risk management, explainability is heavily embedded in regulations like CRR. Credit risk management forces 'transparent by design' models, therefore, explainability is less of an issue. Regulations/supervisory authorities are slow to change on credit risk, possibly to the more international nature and societal importance of regulation in this use case. Changing these regulations to allow for AI-models that are more complex will be an incremental process that takes time and trust in the safety of such models.

3.3 Anti-money laundering

Two banks as well as DNB were interviewed with regard to AML. The use case of one bank involves an AI-solution developed to detect fraudulent activity in corresponding banking transactions. The AI-solution consists of two algorithms (models): a deduplication algorithm and a classification algorithm, which are used to detect possible fraudulent transactions. AML investigators check these transactions. Therefore, there is a human-in-the-loop. The AML investigator receives explanations (e.g. the most important features) as part of the outcome (whether a transaction is flagged as suspicious) of the AML model. In addition to the outcome of the AI model, various information sources are used by the AML investigator to further investigate and determine fraudulent activity. Explanation, in a broader sense, to other stakeholders is via (technical) documentation and various internal processes.

The use case of the other bank concerns machine learning used for transaction monitoring. In the past, transaction monitoring was only rule-based. Currently, a number of machine learning (ML) models are used in conjunction with a rule-based methodology. For instance, there is a supervised ML model that is used as noise reduction on the output of the (business) rule-based system. Furthermore, there is also a supervised model that gives customers scores based on suspicion of money laundering practices. In addition, there is an unsupervised anomaly detection ML model. The output of the models is intended for transaction monitoring analysts (in total about 350 analysts). These have expertise in recognizing integrity risks, but these analysts are generally not concerned with assessing the quality of model output, which is done by a quality assurance analysts.

The ML model output includes extensive information (which can be considered explanation) about suspicious situations, e.g. indicate the most relevant features, as opposed to rule-based systems. This explainability aspect of these (modern ML) models is thus an important part of the subsequent analysis done by the analyst.

This analyst uses a multitude of other data (sources) outside the detection models for further verification. The role of the analysis entails that also in this use case, there is a human-in-the-loop, who is the most important stakeholder in need of explanations.

The application of AI in the AML process is bearing fruit, as results are improving compared to traditional models: fewer false positives and fewer false negatives (so more suspicious transactions are reported). Both internally for banks, but also for supervisory authorities a change of mindset is often still required to transition from the traditional way of thinking in thresholds (encoded in business rules), to more probabilistic thinking about the features of an AML case (locked up in modern machine learning methods). With the latter, explanations can be more complex, but not of less quality.

DNB ensures that banks comply with the Anti-Money Laundering and Anti-Terrorist Financing Act. Currently, DNB does not impose any requirements on which AI model is used for AML, as long as it can be explained, both to the supervisory authority and internally. Exactly what sufficient explanation is for which type of model in various contexts is not defined by DNB, neither in specific requirements (which perhaps is not desirable due to the highly varying contexts in which AI is used), nor general guidelines. For the time being, there is also no framework in which explainability is defined, which is directly applicable to this use case. In the context of controlled business operations, a bank must be able to explain how its systems work. If a bank cannot explain an AI application, both to the supervisory authority and internally, there may be uncontrolled business operations with regard to this specific part and that the bank does not sufficiently manage its risks. Figure 4.3 shows the types of explanations relevant for the supervisor (JST) according to DNB and the banks.

Figure 4.3 Types of explanation relevant for the supervisor according to the supervisor and the particpating banks

Type of explanation	Relevant for supervisor (JST)	
	According to DNB	According to particpating banks
The reasons, details or underlying causes of a particular outcome	■	■ / -
The data and features used as input to determine a particular outcome	■	■ / -
The data used to train and test the AI system	■	■ / -
The performance and accuracy of the AI solution	■	■
The principles, rules, and guidelines used to design and develop the AI solution	■	■
The process that was used to design, develop, and test the AI solution	■	■
The process of how feedback is processed	■	■
The process of how explainers are trained	■	■ / -
The persons involved in design, development and implementation of AI solution	■	■ / -
The persons accountable for development and use of the AI solution	■	■ / -

In summary, AML is one of the use cases that can benefit most from AI in terms of improving results. So far, the issue of explainability did not hinder the use of more complex AI/ML models. The internal AML analyst/investigator (human-in-the-loop) is viewed as the most important stakeholder with regard to explanations, and additionally, this investigator is trained to work with and understand model output.

4 Conclusions and way forward

This chapter discusses the mutual expectations of supervisory authorities and banks on the topic of xAI, draws conclusions on the usability of the conceptual framework, and presents recommendations on how to proceed.

4.1 Mutual expectations

One of the main findings of this study is that the explainable AI is high on the agenda of banks and supervisory authorities. Within banks, it is or will become part of an ethical framework. Such a framework generally builds on existing principles or procedures, but there is a trend towards more unification of principles and a more explicit focus on AI. From the perspective of supervisory authorities, explainability is not exclusively an ethical concern, as it is also relevant from a prudential and legal perspective (e.g. a prudential or legal framework such as CRR, lending standards, Wwft, and GDPR, and contain requirements that include explainability). This broad scope involves a broad range of actors, including data protection supervisory authorities.

The use of complex AI-models by banks is increasing although still limited.
Reasons for the slow adoption are:

1. Time needed to become familiar with and implement complex models and especially xAI-methods. Banks have accumulated considerable experience in terms of Machine Learning (ML), and aim to continuously develop their knowledge on ML. However, the xAI field is still developing rapidly. Deciding what xAI framework to choose (SHAP, etc.) and how to implement it, is often a difficult process as in a short period new methods might make a current xAI-method obsolete.
2. Uncertainty as to whether financial regulations (like lending standards, WFT, Wwft, CRR etc.) or the supervisory authorities would allow it.
3. For many use cases, traditional models are sometimes adequate.
4. Internal hesitation to implement complex AI models in customer facing applications.
5. AI models that are more complex are difficult to maintain and monitor over time (apart from development).

What do banks expect from supervisory authorities?

1. Supervisory authorities current approach is not an obstacle for banks to develop and implement AI solutions. The regulatory framework in many use cases is sufficiently clear as not to obstruct financial entities implementing AI into their operations. For the near future, setting shared 'rules of the game', i.e. joint guidance, is considered desirable. The underlying assumption here is that xAI will become a topic of increasing regulatory and legislative interest. So rather than top-down regulation, a shared vision (regulatory clarity) on what proper (x)AI is would be beneficial; principles are welcomed, as opposed to strict rules. Clear guiding principles would have the added benefit of leading to a level playing field. Banks and supervisory authorities can attain common ground as they uphold the same principles.
2. Banks expect supervisory authorities to increase their knowledge on (x)AI. E.g., supervisory authorities should understand the limitations of xAI. xAI-methods can be helpful in providing explanations, but generally only demonstrate correlations. Supervisory authorities also must be aware that xAI-methods cannot and do not make all AI-models fully explainable and that full explainability/ transparency might lead to gaming in some use cases.
3. Banks call for more cooperation (common vision/opinion) between supervisory authorities and alignment of supervisory authorities on a national and EU level. In addition, the regulatory divergence, which occurs due to different interests (e.g. financial versus privacy interests of different supervisory authorities), should be addressed.

4. Various ways to increase bank-supervisory authorities contact were mentioned including a) a point of contact where banks can safely discuss potential AI innovations, b) a sandbox-like environment where technical developments can safely be shared, and c) a platform to safely discuss the ethical dilemmas that banks might face.
5. Banks are of the opinion that the degree of explainability highly depends on whether there is a human-in-the-loop. Therefore, the question is to what extent a human-in-the-loop can alleviate the strictness/requirements for explainability, as opposed to a (fully) automated process based on AI/ML.
6. According to the banks, the supervisory authorities perspective is less pronounced. Perhaps this is the case because they are less aware of the cutting edge of AI development. However, more detail on their stance on xAI would be very helpful for banks.

What do supervisory authorities expect from banks?

1. The interviewed supervisory authorities will most likely not restrict the use of specific AI-models by banks beforehand. However, according to the interviewees from DNB and AFM, AI models should be explainable, either intrinsically or with a post-hoc method. The required level of explainability depends on the risks the AI model poses for the consumer, the transparency and stability of the financial market and applicable legal frameworks. In other words, the required level of explainability highly depends on the use case and the stakeholders.
2. Whether a certain post-hoc explanation is deemed sufficient also depends on the use case and stakeholders.
3. The matrices (see chapter 4) based on the framework indicate that the interviewees from DNB and AFM seem to adopt a broader scope of explainability, i.e. the types of explanation required by a supervisory authorities, compared to banks. This difference in perspective primarily concerns the question whether specific informational needs from the supervisor should be regarded as proper explainability concerns or as instances of regular supervisory information requirements that are not specific for AI.

4.2 Usability of the conceptual xAI framework

The conceptual framework was reported to be useful in mapping mutual expectations of stakeholders regarding types of explanations. The framework can thus serve as a starting point to discuss which explanations should be provided and how to provide these. The framework should not be regarded as a checklist to prescribe how explainability should be implemented. However, the framework provided a context to move forward jointly in this specific project.

The matrices in chapter 4 indicate that the interviewees from DNB and AFM require all types of explanations. This could be interpreted as micromanagement, according to the participating banks. The question is what process or guidance is useful so that the supervisory authority can trust that the use of x(AI) is up to standards.

The conceptual framework is based on the assumption that explanations are contextual and depend on the specific use case, their stakeholders, and the stakeholders' interests. This assumption has been confirmed in the project. However, it was argued that explainability must be considered together with aspects like fairness, bias, privacy, security et cetera. For example, in specific cases explanations cannot be disclosed because of security reasons. Occasionally, these conflicting aspects create dilemma's, e.g. in certain cases bias might be unavoidable.

4.3 Key takeaways and way forward

One of the objectives of this study was to exchange ideas on and further develop a shared conceptual framework on xAI by applying it to three practical use cases. Furthermore, the aim was to facilitate the discussion between supervisory authorities DNB and AFM and the participating banks regarding the application of AI models, mutual expectations, and the types and degrees of desired explainability for a responsible use of AI in the financial sector. Beside the findings on the usability of the conceptual xAI framework, we also want to highlight the following key takeaways of this preliminary study found in the facilitated discussion on xAI:

- First, based on the interview sample, there appears to be a disparity between the interpretation of the scope of explainable AI from the supervisory authorities as compared to interpretation of the participating banks, based on the aggregated individual viewpoints recorded in this study. DNB and AFM interviewees indicated they require all types of explanations covered by the framework, while bank interviewees only consider a subset to be relevant instances of explainability concerns for each use case. Therefore, looking forward it could be beneficial for banks and supervisory authorities to formulate a more nuanced and detailed viewpoint on the desired scope of xAI for different types of AI models. Developing such nuanced and detailed viewpoints can also provide clarification regarding the minimal required level and type of xAI per use case, and facilitate considerations regarding the trade-off between performance and explainability.
- Secondly, there is a need for enhanced cooperation between supervisory authorities, and alignment of supervisory authorities on a national and EU level on the topic of AI and explainability of AI models. Developments in AI specifically call for increased cooperation between supervisory authorities due to AI's impact on both the technical and social domain. Looking ahead, it would be useful to discuss and create shared principles regarding explainable AI amongst supervisory authorities on both national (for instance, DNB, AFM and AP) and EU level (for instance EBA, ECB and SSM).
- Thirdly, the need to increase contact between bank and supervisory authorities to safely discuss and share AI innovations was brought up and was found to be of importance. Looking forward, it would important to discuss how and via which communication channels the contact between banks and supervisory authorities (both national- and EU level) on the topic of AI can be further facilitated.
- Fourthly, explanations were found to be highly contextual and to vary per use case. The framework as used in this study was indicated in the interviews to be a potentially valuable starting point, but not as a comprehensive approach, to consider xAI for use cases. Looking forward, to determine the minimum desired level and type of explanation for a use case, a more granular framework of types of explanations could be useful. This might be achieved by building on a set of AI principles (such as proposed in the first point) and creating a more granular framework which considers use cases and model choice within those use cases to give a practical summation of xAI aspects to consider.

The challenges presented by these key takeaways will be taken up by the iForum in 2021.

References

ACPR. Discussion paper "Governance of Artificial Intelligence in Finance"; June 2020.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

European Commission (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence.

Gerlings, J., Shollo, A., & Constantiou, I. (2020). Reviewing the Need for Explainable Artificial Intelligence (xAI). *arXiv preprint arXiv:2012.01007*. Gerlings, J., Shollo, A., & Constantiou, I. (2020). Reviewing the Need for Explainable Artificial Intelligence (xAI). 11.

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31-57. Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57.

Lundberg, S., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*. Lundberg, S., & Lee, S. I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1-38.Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. Van den Berg, M., & Kuiper, O.X. (2020). Explainable AI in the Financial Sector Van den Berg Kuiper septemberSeptember 2020.pdf.

The UK Information Commissioner's Office / Alan Turing Institute. "Explaining Decisions made with AI"; May 2020.

Van den Berg, M., & Kuiper, O. (2020). XAI in the Financial Sector.