A CONCEPTUAL FRAMEWORK FOR EXPLAINABLE AI (XAI)

XAI IN THE FINANCIAL SECTOR

HOGESCHOOL UTRECHT

IMPACT YOUR FUTURE

TABLE OF CONTENTS

ABSTRACT

The use of AI is on the rise in the financial sector. Utilizing machine learning algorithms to make decisions and predictions based on the available data can be highly valuable. AI offers benefits to both financial service providers and its customers by improving service and reducing costs. Examples of AI use cases in the financial sector are: identity verification in client onboarding, transaction data analysis, fraud detection in claims management, anti-money laundering monitoring, price differentiation in car insurance, automated analysis of legal documents, and the processing of loan applications.

With the increasing usage of AI there is a call for it to remain understandable and transparent. Some machine learning algorithms have become so complex that it becomes more and more difficult to explain how a certain decision or prediction is reached based on the data. Especially in the case of AI techniques such as deep neural networks the process from input to output is virtually impossible to interpret. Explainable AI (abbreviated to XAI) aims to provide a solution to this 'black box' problem. Such a solution is a prerequisite for a large scale deployment of AI in the financial sector. Compared to other industries the financial sector is held to higher societal standards concerning trust and transparency. For example, the responsible use of personal data is a major factor for trust in financial institutions. XAI is an important tool to increase trust in the use of AI by the financial sector. Now the question is how to effectively deploy XAI in the financial sector.

Together with experts from the financial sector we studied the field of XAI and developed a framework to analyze XAI in finance. The framework is aimed at identifying what type of explanation different stakeholders in the financial sector require, both in terms of the required information and in terms of effectively conveying that information. We defined XAI as follows: 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. As such, the importance of XAI extends beyond offering developers insights in the AI algorithms, and should be viewed as an essential tool or method for providing the required level of understanding for all stakeholders, internal or external to an organization. Once it is clear what types of explanation are required in a given use case, appropriate methods and techniques can be applied to provide these explanations.
Considering the speed at which AI develops, a clear framework of stakeholders' requirements can be a crucial tool for financial service providers, regulators, and policy/law makers to regulate and stimulate the use of AI. Furthermore, a framework cognizant of the capacities of the various stakeholders of the AI system ensures a human-centric focus, which is essential in designing proper AI in conjunction with XAI.

Dr. Martin van den Berg
martin.m.vandenberg@hu.nl
Dr. Ouren Kuiper
ouren.kuiper@hu.nl

Hogeschool Utrecht, Lectoraat Artificial Intelligence
Version 1.1 - November 2020

1. INTRODUCTION

The use of artificial intelligence (AI) is rapidly increasing in the financial sector [Ryll et al.; NVB; McWaters]. Utilizing machine learning (ML) algorithms to make decisions and predictions based on the available data can be highly valuable, both for financial service providers and its customers, by offering better service and saving costs. However, especially as the mathematical models become increasingly complex, drawbacks to this technology might arise as transparency is lost. In the last several years, the topic of explainable AI (XAI) has gained increased attention in AI research, as an attempt to counteract these downsides. In this research we will explore the main developments on the subject of XAI and how these relate to the financial sector specifically.

We conducted a literature study on XAI to explore the main developments. Based on the literature study we subsequently defined XAI and developed a conceptual framework. This framework will be used as a starting point to study the impact of XAI on the financial sector. However, the framework can also be used in other industries.

We discussed the framework with a focus group with participants from the financial sector and our AI research group at the Hogeschool Utrecht. They provided us with valuable feedback. Many thanks to Joost van der Burgt, Tom Leenders, Stefan Leijnen, and Sieuwert van Otterloo.

This paper is structured as follows. In section 2 we first define XAI. In section 3 we discuss one of the key concepts in XAI, namely the stakeholder. Since XAI is meant to provide proper or fitting explanations for each stakeholder, we examine the purpose of explanations in section 4 and the traits and qualities of explanations in section 5. In section 6 we present the conceptual framework, which aim is to provide insight which type of explanation to provide to which type of stakeholder. In section 7 we discuss methods to provide an XAI explanation given various AI techniques. In section 8 we discuss the limitations of XAI. In section 9 we explore in more depth XAI in the financial sector, including an application of the conceptual framework in the context of lending. Finally, in section 10 we formulate a conclusion.

2. DEFINITION OF XAI

Explainable AI (XAI), also called interpretable or understandable AI, aims to provide a solution to the 'black box' problem in AI. That is, an AI solution utilizes data (e.g. on an individual's financial situation) and produces an outcome (e.g. rejecting a certain loan). However, in this process there is generally no output that explains how or why the outcome is reached based on the data. Especially in the case of AI techniques such as deep neural networks, the process from input to output is virtually impossible to interpret even with knowledge of the inner workings, weights, and biases of the system. XAI explains why or how the AI solution arrived at a specific decision.

Several definitions of XAI have been proposed, e.g.:

- "Explainable AI is a set of capabilities that describes a model, highlights its strengths and weaknesses, predicts its likely behavior, and identifies any potential biases. It can articulate the decisions of a descriptive, predictive or prescriptive model to enable accuracy, fairness, accountability, stability and transparency in algorithmic decision making" [Sicular et al.].
- "Given an audience, an explainable AI is one that produces details or reasons to make the functioning clear or easy to understand" [Arrieta et al.].
- "Explainable AI (XAI) refers to methods and techniques in the application of AI technology such that the results of the solution can be understood by human experts" [Wikipedia].
- "XAI refers to an explanatory agent revealing underlying causes to its or another agents' decision-making" [Miller].
- "Explainability is the ability to explain the reasoning behind a particular decision, classification or forecast" [Dwivedi et al.].

Definitions of concepts surrounding XAI

To fully understand XAI and its goals, it should be understood what it means for an explanation to be 'interpretable' or 'understandable'. Several authors on XAI (e.g. Lipton; Mittelstadt et al.) emphasize that XAI is often hastily defined without proper understanding of the perspective of all parties involved. Thus, first we will explore several terms crucial in the understanding and definition of XAI: transparency, interpretability, and explanation. Furthermore, definitions of understandability, traceability, and auditability warrant examination as these concepts are especially relevant to AI in the financial sector.

Transparency

The term transparency is used in two ways with regard to XAI.

1. Transparency of a system or model is the property to be understood by a human. Most authors on XAI assume that this understanding is direct and the system as-is is interpretable, i.e. without further requirements. Transparency in this usage has been defined as:
   - o "A model is considered to be transparent if by itself it is understandable." [Arrieta et al.]
   - o "The opposite of black-box-ness is transparency, i.e., the search for a direct understanding of the mechanism by which a model works." [Arrieta et al.]
   - o "Interpretable systems [are systems] where a user cannot only see, but also study and understand how inputs are mathematically mapped to outputs. This implies model transparency [..]" [Doran et al.]
2. Transparency of the implementation of the AI solution is about not (unintentionally) concealing information for stakeholders, such as customers or auditors, but rather attaining openness:
   - o "Transparency is about being clear, open and honest with people about how and why you use their personal data." [ICO] The ICO report has a more user-focused (rather than model-focused) view on XAI, and their usage of the term transparency reflects that.
   - o The EU High-level expert group on AI defines these three elements of transparency: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system [HLEG].
     - » Traceability, which pertains to "the data sets and the processes that yield the AI system's decision." Traceability is a requirement for auditability as well as explainability.
     - » Explainability "concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system)."
     - » Communication entails that AI should not pose as human. Additionally, "the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand."

The two ways in which the term 'transparency' is used are connected, as an AI solution that is very easily interpretable is more easily implemented in an open and communicative fashion. However, an AI solution using an easily interpretable model can still be presented in a non-transparent way, for instance by not informing users that an AI is used at all. In the context of AI in the financial sector the second kind of transparency, relating to openness and truthfulness, is the focus of this paper.

Interpretability

A term closely related to transparency of a system or model is interpretability. Interpretability has been defined as:

- "The ability to explain or to provide the meaning in understandable terms to a human." [Arrieta et al.]
- "To which extent the model and/or its predictions are human understandable." [Guidotti et al.]
- "Systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation." [Biran & Cotton]
- "Interpretability refers to the concept of comprehensibility, explainability, or understandability. When an element of an AI system is interpretable, this means that it is possible at least for an external observer to understand it and find its meaning." [HLEG]

In XAI transparency and interpretability are sometimes used interchangeably. For transparency of a system or model these terms are indeed interchangeable, however, in the use of 'transparency' as openness and truthfulness these terms have a different nuance.

Explanation

While some systems are transparent by nature, an explanation can be provided to make a system or its outcome understandable. Explainability entails that an explanation can be formulated. In the literature several aspects of what constitutes an explanation of an AI solution are consistently reported. Namely an explanation:

- is highly contextual, as different people in different settings may require different kinds of explanations;
- should be able to serve different purposes;
- has different stakeholders (such as staff, end-users, auditors, supervisors);
- can have different formats (such as text, visuals);
- can be generated in different ways (with different methods and techniques);
- is not always possible (is at the expense of accuracy or model performance);
- is not always necessary (in case the concerns of stakeholders do not require an explanation);
- requires a process in which the explanation is provided;
- requires a selection of information to be part of an explanation;
- has different levels of goodness or quality (explained further in section 5).

Understandability, traceability, and auditability

For this paper, we generally hold that if an AI solution is transparent, it is interpretable. Interpretability, in turn, entails that it is understandable. Understandability in turn we hold to be fairly self-evident: comprehensible and able to be understood. As this string of definitions shows, exact definitions of a term can be troublesome and often rely on other terms or even context [Lipton; Arya et al.]. This difficulty can be partially alleviated by understanding how an (X)AI solution effects various specific parties or stakeholders; these are explained in detail later in this document.

Traceability, pertains to "the data sets and the processes that yield the AI system's decision." [HLEG]. Traceable entails it can be determined which data is used, and by what process the outcome of the AI solution is reached. It is especially important in relation to auditability, the trait of being capable of being audited. Traceability is especially important in medical, financial, and other domains with strong legislation and potential ethical risks. For other domains traceability might be less important, e.g. an AI solution that detects unripe fruit to remove from a conveyor belt which does not directly impact humans.

DEFINITION OF XAI IN THIS PAPER

We combine some of the earlier mentioned definitions of XAI into this new definition:

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 highlights that:

- An explanation has one or more stakeholders and every stakeholder may require a different kind of explanation according to his/her concerns.
- An AI solution can have multiple stakeholders requiring an explanation.
- XAI is a capability, i.e. the ability to provide a meaningful explanation. In the context of XAI this capability not only consists of methods and techniques, but also of processes (social interaction) and people (stakeholders).
- XAI produces an explanation in the form of details, reasons or underlying causes. This information needs to be tailored to the stakeholder so that he/she can understand it and addresses his/her concerns.
- XAI is about the functioning and results of an AI solution. XAI is thus meaningful in the context of AI.

In some of the literature on XAI, the aspect of the stakeholder is not explicitly considered (e.g. [Adadi & Berrada; Guidotti et al.]), while in others it has a central role (e.g. [Arya et al.; Bracke et al.; ICO]). However, most authors explicitly acknowledge that an explanation is context-dependent; the stakeholder receiving explanation can be seen as the context. We argue that, especially in the domain of AI in the financial sector, identifying the stakeholders and their respective needs (and rights) is a crucial aspect of XAI. Thus, in our definition XAI has a wider impact than exclusively offering model developers insights in the AI algorithms, as it should also effectuate the required level of understanding for all stakeholders, both internal or external to an organization

Visual overview of XAI

We constructed a visual overview of XAI in the context of an AI solution as can be seen in figure 1.

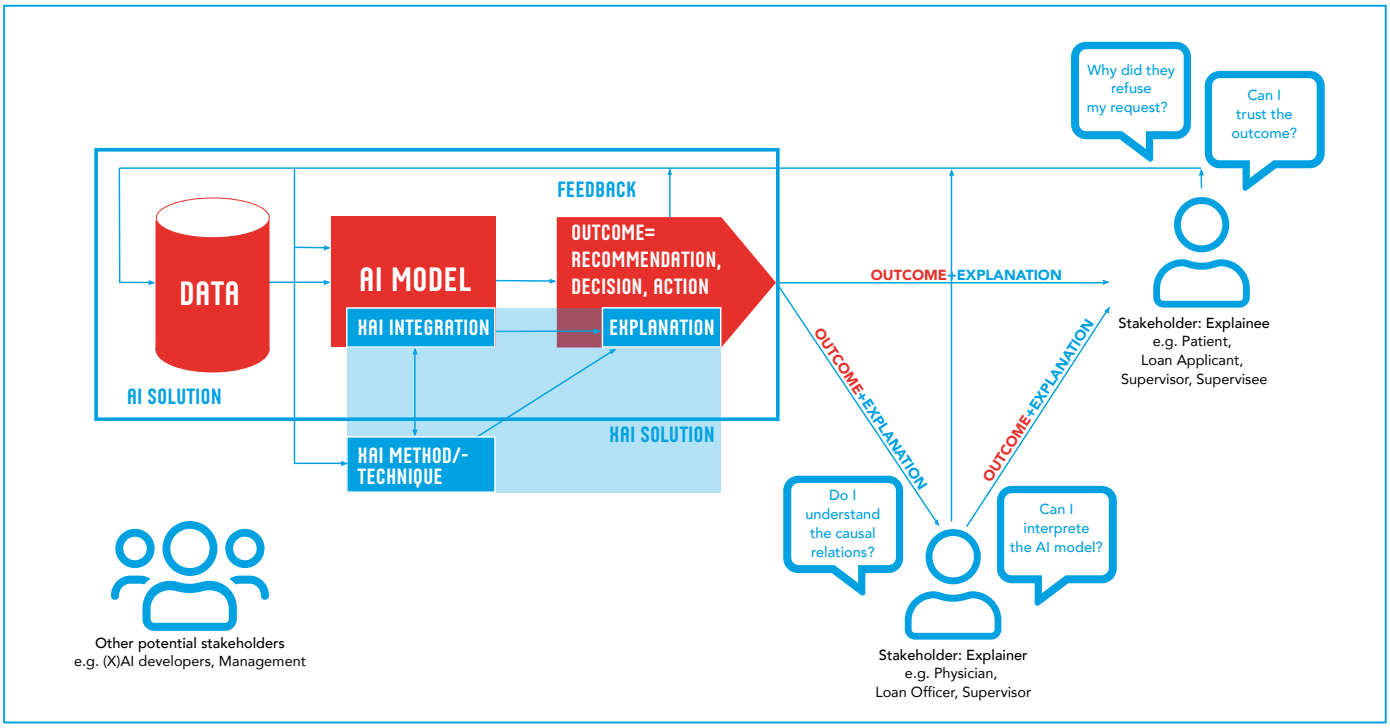


Figure 1. An XAI solution in the context of an AI solution with stakeholders.

In figure 1, the XAI solution (blue squares) adds an explanation to the outcome (e.g. the rejection of a loan) of the AI solution. In this example, the explanation would make it understandable why a loan was rejected. Stakeholders can have a special relation in that one can be an explainee (e.g. a loan applicant or patient) while one can be an explainer (e.g. loan officer or a doctor). The explainer may use the outcome and an explanation to give their own explanation to the explainee to aid them in understanding the outcome. Various stakeholders might be either an explainer or an explainee depending on context; virtually always different stakeholders require specific explanations to satisfy their specific concerns. In figure 1, interpretability (i.e. model transparency) is an aspect of the AI model. This is of importance to for instance the developer of the XAI solution. Transparency of the whole system (i.e. openness) is a trait of the (X)AI solution as a whole and is primarily of importance to stakeholders such as patients or supervisors. Traceability here would mean that the outcome provided by the AI solution can be logically found to be a result of the data and the AI model.

An AI system can be so complex that an XAI method/technique might require full integration in the main AI model, rather than being a modular addition needing only the information relevant for an explanation. How various AI techniques and differing contexts require different explanation methods is further explored in the following sections. Regardless of implementation, the XAI method/technique must be able to provide details, reasons and causes to be used to (automatically) provide an explanation to a stakeholder, possibly via a human in the loop (i.e. an explainer).

Virtually every implementation of AI is bound to have a stakeholder that needs to be informed of the operation of the AI solution. At the very least to check the AI's performance, but generally also to justify usage of the AI solution. In this sense, virtually all use of AI requires accompanying XAI in some form. In the case of rudimentary applications of AI, the XAI solution can be said to be –rather than an automated process– an expert with sufficient knowledge of the AI system to construct and convey explanations to the required stakeholders. For instance, an AI developer can explain to higher management what biases a loaning algorithm might have. However, generally when the term XAI is used, it refers to an XAI solution that is itself a software application that generates explanations automatically (albeit possibly still intended to be communicated further by human explainers). More complex AI solutions, i.e. lacking model transparency, require specific automated XAI methods to be explainable (see section 7 for various existing XAI methods depending on the AI model used). Importantly, a good XAI solution at minimum has been thoughtfully designed and clearly formalized, irrespective of its implementation.

3. STAKEHOLDERS

An explanation is always dependent on context: its exact form depends on what information is required (i.e. what concerns should be met) and on the capacities of who is receiving the information (e.g. a child requires a different explanation as compared to a domain expert [Guidotti et al.]). The individual or party requiring an explanation is referred to as the stakeholder.

Stakeholders can take two different roles in regard to XAI: the explainee and the explainer.

- The explainee is the end-user of the outcome or result of the AI solution, whether it is a recommendation, decision or action, e.g. a loan applicant or a patient. The explainee can also be a party that is representing the concerns of a group of end-users, e.g. a supervisor, regulator, or an organization that defends the interests of consumers. Typical concerns of an explainee are: "Can I trust the outcome?" or "Why did they refuse my request for a loan?".
- The explainer can be the one who provides the explanation to the explainee, e.g. a loan officer or a physician. It is a person who is interested in the overall functioning of the AI solution with concerns such as "Can I interpret the AI model" or "Do I understand the causal relations".

Different explainer-explainee relationships exist, like the physician-patient, loan officer-loan applicant, supervisor-supervisee, or domain expert-supervisor. Being cognizant of these relations is important in developing the right types of XAI, as the information that the explainer requires to successfully aid the explainee in understanding is different from what either party would need in isolation to understand the outcome. A capable explainer and a high quality XAI solution will outperform the sum of their parts ('human-AI symbiosis').

If a company employs an AI solution the staff are generally explainers [ICO]. They need to relay meaningful information on the outcome of the AI solution. On the other hand, there are two types of explainees in a professional context: end-users and auditors. The end-users, i.e. individuals affected by the decisions of the AI solution, are generally customers. Auditors (e.g. regulators, external or internal reviewers) are explainees as they are charged with monitoring or overseeing the production, deployment and use of the AI solution.

Depending on the issue a stakeholder can have different relations to the AI solution, and thus these roles can shift in a different context. For instance, an AI developer can be explainee when using XAI information to improve the system, but that developer can be explainer to others in the organization. An auditor might receive explanation from an explainer when auditing an organization but be the explainer when relaying the information further.

In the financial context, Bracke and colleagues identify at least six different types of stakeholders [Bracke et al.]:

- Developers, i.e. those developing or implementing an AI application;
- First line model checkers, i.e. those directly responsible for making sure model development is of sufficient quality;
- Management responsible for the application;
- Second line model checkers, i.e. staff that, as part of a firm's control functions, independently check the quality of model development and deployment;
- Conduct regulators that take an interest in deployed models being in line with conduct rules.
- Prudential regulators that take an interest in deployed models being in line with prudential requirements.

We propose the following list of stakeholders:

- End user (e.g. a customer)
- Explainer to the end user
  - External advisor (e.g. a financial advisor)
  - Internal advisor (e.g. a loan officer)
- AI developer
- Domain expert
- Executive management
- Management responsible for the AI solution (1st line)
- Operational control (2nd line)
- Audit (3rd line)
- Other stakeholders (e.g. regulators and auditors)

This list captures the most important stakeholders in the context of a general AI solution in the present day. The other stakeholders are domain-dependent and should be determined depending on the exact AI solution. For instance, in the financial sector regulators are stakeholders as they are tasked to oversee whether rules and regulations are met. The above list also incorporates the 'three lines of defense' principle [IIA], especially relevant in the financial sector. Furthermore, in an Agile/Scrum environment, the product owner could be

comparable to what we label domain expert, while the scrum-master might have similar interests to that of the AI developer in terms of explanation. Note, that one person can fulfil several types of stakeholder roles. E.g., a domain expert can also be an explainer for the end user.

Society as a whole, or even future humans can also be regarded as a stakeholder of AI solutions. While the current applications of AI are highly specialized and the risks are not as great what some believe the risks of potential 'general AI' are [Stephen Hawking, BBC], it is prudent to be aware of potential unforeseen consequences of the automation and risk of bias that comes with AI.

4. PURPOSE OF AN EXPLANATION

Providing an explanation can serve various purposes. The precise purpose depends on the context and stakeholders of the XAI. When, for example, personal data are used the GDPR stipulates that a user has the right to an explanation. Legal compliance is thus a purpose. A general purpose of an explanation in the context of AI is to increase users' trust and confidence. Other, more general purposes are to justify, verify and improve decisions [ICO]. Reasons to verify decisions are e.g. whether the decisions are fair, ethical, or accurate. Considering the reliability and robustness of the AI solution can also be purposes to want an explanation.

Different stakeholders can have different goals that an explanation can meet. Developers of an AI solution might want to increase their system's transparency to be able to improve it. Governments and regulators might want to have insight in AI solutions to be aware of the risks for citizens. End users such as consumers might want to be able to trust AI solutions, or better understand what it can do to better utilize it.

The following lists of possible purposes have been proposed:

- Trustworthiness, causality, transferability, informativeness, confidence, fairness, accessibility, interactivity [Arrieta et al.].
- Interpretability, accuracy, fidelity, fairness, privacy, usability, monotonicity, reliability/robustness, causality, scalability, generality [Guidotti et al.].
- User benefit, societal acceptance, regulatory and compliance, system development, owner benefit [NIST].

Two main types of goals of explanations can be differentiated. Firstly, some goals are instrumental or functional in that they can improve the main function of the AI solution and thereby benefit a person or party (both owner and end user of the AI solution). AI experts, such as AI engineers and AI developers might seek these kinds of explanations. XAI can aid in the better understanding of the AI techniques by informing us of causality (beyond only offering correct inferences also an understanding of the underlying phenomena causing a correlation between input and output) [Hagras]. End users might want to understand their specific case. XAI can help probe the mechanisms of ML systems (e.g. we seek an interpretation of how the system works) and to relate explanations to particular inputs and examples (e.g. we want to comprehend how an input was mapped to an output) [Doran et al.]. Insight in AI solutions offered by XAI can even help discover new abilities of AI [Adadi & Berrada].
Secondly, there are goals of a more ethical, social, and legal nature. These second type of goals might conflict with the first, as regulators and governments might require transparency of AI solutions that do not increase the performance of these solutions. The social explanation is used to enhance the explainee to appropriate trust and use, and avoid distrust and misuse [Cui et al.]. These goals are often demonstrating qualities like trustworthiness, fairness, accessibility, or privacy awareness [Arrieta et al.]. In general, justification and trust are central in this type of goal for XAI.

The following list of purposes is stated in the ICO report [ICO], which focuses on AI in a business-customer relationship:

- Legal compliance, i.e., comply with law (like GDPR). Explaining AI-assisted decisions to those affected will help to give better assurance of legal compliance, mitigating the risks associated with non-compliance.
- Build trust with customers. Explaining AI-assisted decisions to affected stakeholders makes good business sense. This will help to empower stakeholders to better understand the process and allow them to challenge and seek recourse where necessary.
- Improve internal governance. Explainability provides oversight of what AI solutions do and why.
- Informed public. As more organizations incorporate explanations to individuals as a core element of their AI-assisted decision-making systems, the general public will gain an increasing awareness of when and where such decisions are made.
- Better outcomes. Explanations support more consistency and fairness in the outcomes for different groups across society.
- Human flourishing. Giving stakeholders explanations of AI-assisted decisions helps to ensure that the use of AI is human-centric.

A company not explaining AI-assisted decisions could face regulatory action, reputational damage, and disengagement by the public [ICO]. Therefore, properly implementing XAI in conjunction with AI-assisted decisions is not only an improvement for such AI solutions, but a necessity. Additionally, XAI should not be a purely epistemological endeavor, i.e. knowledge for knowledge sake, but it should be relevant and address the stakeholders' concerns and actually aid them in taking informed action.

5. TRAITS AND QUALITIES OF A PROPER EXPLANATION

The core concept of XAI is the explanation. The Cambridge dictionary defines an explanation as: "the details or reasons that someone gives to make something clear or easy to understand". For this paper, we will not fully explore the nature of what constitutes an explanation, however, it is important for XAI to understand what constitutes a proper explanation depending on context and stakeholder. Coming home to find a broken glass next to a table, the mention of 'cat' might be enough explanation of what happened (if one has a cat); conversely, an extensive explanation including descriptions of the force of gravity, the surface tension of glass, and the molecular structure of the flooring is factually a more complete explanation, but not the required one. A good or 'proper' explanation is not only about providing all the information, but to do so in a manner that leads to stakeholder understanding.

Explanations can have several aspects that make them work. Explanations can be contrastive, as particular counterfactual cases are especially informative. That is, it is often more interesting to explain why event P happened instead of event Q, than it is to explain why event P happened in isolation. Explanations are always selective: explanations are not a complete description of the casual chain leading up to an event. Humans are good at identifying one or two causes from an infinite space of possibilities to be the 'right' or most informative explanation [Miller].

Furthermore, explanations are social. In the transfer of knowledge, assumptions about the stakeholder's prior knowledge and views influence what constitutes a proper explanation [Miller]. There is always a mental model on the other party in an explanation [Gunning], especially relevant to explainer-explainee relations. Some authors argue that probabilities do not matter in terms of explanation [Miller]. Probabilities do not inform us of causal relations, only correlation, and in that sense, they do not explain phenomena. However, for garnering trust, probabilities are certainly valued. For example, most people do not know how a jet engine works, yet fully trust commercial airplanes based on their overwhelming statistical safety.

Various types of explanations can be distinguished. A first division that can be made is that of a: process-based or outcome-based explanation [ICO; Miller]. A process-based explanation gives information on the governance of the AI solution across its design and deployment; the explanation is about "the how". An outcome-based explanation tells what happened in the case of a particular decision; the explanation is about "the what".

Stakeholders affected by a decision (explainees such as loan applicants and patients) are most likely more interested in an outcome-based explanation. Stakeholders in the role of explainer, such as loan officers and physicians, are probably interested in both process- and outcome-based explanations. Auditors and supervisors are most likely interested in process-based explanations. "How" explanations are useful for AI-developers, while "why" explanations are useful for end-users [Dwivedi et al.].

A proper explanation should contain the right amount of detail, and it should reveal the boundary conditions of a model [Mueller et al.]. In addition, aspects of an AI solution can be globally interpretable or locally interpretable and explanations can be global or local [Adadi & Berrada; Guidotti et al.; Mueller et al.]. That is, a global explanation reveals the inner workings of the entire AI solution (potentially including a case at hand), a local explanation offers insight in a specific outcome.

Depending on the stakeholders' concerns, a different type of explanation is required. The ICO-report identifies six types of explanations:

- Rationale explanation: the reasons that led to a decision, delivered in an accessible and non-technical way.
- Responsibility explanation: who is involved in the development, management and implementation of an AI solution, and who to contact for a human review of a decision.
- Data explanation: what data has been used in a particular decision and how.
- Fairness explanation: steps taken across the design and implementation of an AI solution to ensure that the decisions it supports are generally unbiased and fair, and whether or not a stakeholder has been treated equitably.
- Safety and performance explanation: steps taken across the design and implementation of an AI solution to maximize the accuracy, reliability, security and robustness of its decisions and behaviours.
- Impact explanation: steps taken across the design and implementation of an AI solution to consider and monitor the impacts that the use of an AI solution and its decisions has or may have on a stakeholder, and on wider society.

The ACPR-report describes four levels of explanations as an attempt to provide a scale for the depth of an explanation. The four levels are: 1) observation: how does the algorithm work (technically-speaking) and what is the algorithm's purpose (functionally-speaking), 2) justification: why does the algorithm produce such a result, 3) approximation: how does the algorithm work (inductive), and 4) replication: how to prove that the algorithm works correctly (demonstrable). An observation type of explanation has the least depth and the replication type has the most depth. Depending on the use case, the context and in particular, the recipients and associated risks, it can be determined which of the four explanation levels is appropriate [ACPR].

Knowledge of the context is essential in determining what kind of explanation should be provided. According to the ICO-report a good explanation is truthful and meaningful, written or presented appropriately, and delivered at the right time. Five context factors are identified [ICO]:

- Domain. This is the sector where the AI solution is deployed. This can affect the explanations stakeholders want. For instance, what stakeholders want to know about AI-assisted decisions made in the criminal justice domain can differ significantly from other domains such as healthcare, finance or gaming.
- Impact on the stakeholder. The 'impact' factor is about the effect an AI-assisted decision can have on a stakeholder and wider society. Varying levels of severity and different types of impact can change what explanations stakeholders will find useful, and the purpose the explanation serves.
- Data used. Data as a contextual factor relates to both the data used to train and test the AI solution, as well as the input data at the point of the decision. The type of data used can influence a stakeholder's willingness to accept or contest an AI-assisted decision, and the actions they take as a result.
- Urgency of the decision. The 'urgency' factor concerns the importance of receiving or acting upon the outcome of an AI-assisted decision within a short timeframe. What stakeholders want to know about a decision can change depending on how little or much time they have to reflect on it.
- Audience it is being presented to. 'Audience' as a contextual factor is about the stakeholders you are explaining an AI-assisted decision to. The groups of stakeholders you make decisions about, and the stakeholders within those groups have an effect on what type of explanations are meaningful or useful for them.

Based on the above views on what constitutes a proper explanation, an XAI explanation may thus contain the following content:

- Outcome of the AI solution:
  - The reasons, details or underlying causes of a particular outcome, both from a local and global perspective.
  - The data and features used as input to determine a particular outcome, both from a local and global perspective.
- Operation of the AI solution:
  - The data used to train and test the AI solution.
  - The performance and accuracy of the AI solution.
  - The principles, rules, and guidelines used to design and develop the AI solution.
- Processes surrounding the AI solution:
  - The process that was used to design, develop and test the AI solution (considering aspects like compliance, fairness, privacy, performance, safety, and impact).
  - The process of how feedback is processed.
  - The process of how explainers are trained.
- Governance in relation to the AI solution:
  - The persons involved in design, development and implementation of the AI solution.
  - The persons accountable for development and use of the AI solution.

	Functional		Ethical, social, legal			
	Understand a specific case	Improve the overall (X)AI solution	Adhere to ethical principles	Build trust	Comply with legislation	Improve internal governance
The reasons, details or underlying causes of a particular outcome.	X	X		X		
The data and features used as input to determine a particular outcome	X	X		X	X	
The data used to train and test the AI solution		X	X	X	X	
The performance and accuracy of the AI solution		X		X		
The principles, rules, and guidelines used to design and develop the AI solution		X	X	X	X	
The process that was used to design, develop and test the AI solution		X	X	X		
The process of how feedback is processed		X	X	X		X
The process of how explainers are trained		X		X		X
The persons involved in design, development and implementation of AI solution			X	X	X	X
The persons accountable for development and use of AI solution				X	X	X

Figure 2. Overview of the general types of content an explanation might have, and the goal that is typically fulfilled by that content. In our framework of XAI, these goals are the addressing of a stakeholder's concern.

In figure 2, we relate the list of possible content of an XAI explanation to the goal it generally has. Thus, addressing a stakeholder's concern, e.g. legal compliance in terms of the correct use of personal data, is the goal of an explanation. The content of the explanation can in turn e.g. be information on 'data used to train and test the AI solution', which fulfils this goal. Note that this overview does not give any insight in the best method to present this information (i.e. the explanation method); this also depends on the AI technique that is used. Explanation methods are explained further in section 8. Depending on the stakeholder, building trust can be dependent on knowledge of a part or, as seen in the figure, the whole (X) AI system. However, generally a single stakeholder will only require a subset of the available information relevant to their concerns in order be able to trust the system.

6. CONCEPTUAL FRAMEWORK

Based on the above knowledge on stakeholders and on the content of explanations given stakeholders' concerns, we constructed a conceptual framework for XAI which can be seen in figure 3. The framework was inspired by that of Bracke et al., but expands both on the type of stakeholders, and the possible explanations. In addition, we give a more concrete overview not only of what type of questions the explanations might answer (i.e. what its goals are), but also what the actual content of such an explanation would most likely be. This gives a more practical framework to help in developing XAI solutions that takes into account the full spectrum of stakeholders and types of explanation at an early stage. The framework also encompasses the processes surrounding the implementation of the XAI solution, rather than only its technical implementation. A broader view on XAI as it is embedded in an organization can facilitate better governance, which is generally both an aim and a requirement of both internal (e.g. operational control) and external stakeholders (e.g. regulators).

In this framework, the end user will often be an explainee and receive information from their advisors (internal of external), i.e. explainers. Generally all three types of stakeholders will want the same type of explanation, as can be seen in the framework. However, depending on the XAI solution, the end user might either receive an explanation directly, or it might be required that an advisor relays the explanation. While an ideal XAI solution might be able to tailor an explanation to any stakeholder, regardless their level of knowledge, a (expert) human in the process will greatly increase the overall capacity to make the end user understand the AI decision.

A further distinction for virtually all applications of AI is that between the (AI) service provider and the external (end) user of the product. The latter is dependent for its information on the supplier or administrator of the AI solution, and generally has a protected legal status, especially in the context of finance. While the framework gives an overview that suits most industries, the exact types of stakeholders may vary by industry and by service.

	Type of stakeholder												
	External		**Service provider**							**Other stakeholder (context)**			
Type of explanation	End user	End user's external advisor	End user's external advisor	AI develo-per	Domain expert	Exe-cutive mgmnt	Opera-tional mgmt (1st line)	Opera-tional control (2nd line)	Audit (3rd line)	Regula-tor A	Regula-tor B	Regula-tor C	.....
The reasons, details of underlying causes of a particular outcome	X	X	X										
The data and featu-res used as input to determine a particular outcome	X	X	X	X									
The data used to train and test the AI solution				X	X	X	X	X	X				
The performance and accuracy of the AI solution				X	X	X							
The principles, rules and guidelines used to design and develop the AI solution				X	X	X	X	X					
The process that was used to design, develop ans teskt the AI solution					X		X	X	X				
The process of how feedback is processed					X		X	X	X				
The process of how explainers are trained					X		X						
The persons involved in design, development and implemantation of AI solution.					X	X	X	X	X				
The pesons accountable for development and use of AI solution						X	X	X	X				

Figure 3. Conceptual XAI framework. The cells with an "X" indicate that this type of content of explanation might be especially relevant for the type of stakeholder.

7. XAI METHODS

Depending on the ML model used in an AI solution, different explanatory methods have been described in the past several decades. Do note, these methods are often constructed to assist the explainer, such as a domain expert, to understand the working of the AI solution and how its decisions are made. Conversely, methods to provide an adequate explanation to an explainee (end user such as a customer) are rare. A typical end user of an AI solution will generally have case-specific ('local') interests, and a full understanding of the AI solution might not be required. The reason most XAI methods focus on the expert explainer, while to our knowledge only a few (e.g. [ICO]) aim at the explainee, might stem from the fact that currently those that are in the field of developing AI solutions also develop their XAI counterparts. In publishing works on XAI, they do so from the perspective of the domain expert. However, as our framework demonstrates, a broader view on stakeholders is necessary in the process of developing XAI.

Questions about the 'why' of AI outcomes "require explanations that are contrastive, selective, and socially interactive" [Mittelstadt et al.]. A call for a focus on philosophy, cognitive science, and the social sciences to aid in XAI has also been made by other authors [Miller]. This again resonates with the above section that the best explanation is not the most complete explanation, but the explanation that answers the stakeholders' concerns.

Transparent and post-hoc explainable techniques

Various authors make the distinction of inherently transparent or interpretable models and post-hoc explainable models [Adadi & Berrada; Arrieta et al.; Guidotti et al.]. The non-transparent models are sometimes also referred to as 'black-box' models. The latter requiring a set of techniques "to reverse engineering the process to provide the needed explanation without altering or even knowing the inner works of the original model" [Adadi & Berrada].

Various types of AI models exist –that are trained by data and give predictive output– using different AI or ML techniques (we will use the term 'technique' instead of methods to avoid confusion with explanation methods). A model might be transparent and easily interpretable based on the nature and implementation of that technique. Examples are a simple –transparent– linear regression model, or a naturally less transparent neural network. In the case of a linear model, the weights given to different features (e.g. age, gender) are easily understood. However, a linear model with highly complex polynomial features might be less interpretable than a simple neural network. Inherently less transparent techniques can be coupled with visualization methods such as saliency maps to increase their explainability [Arya et al.]. The set of these techniques is generally fairly constant in the literature on XAI (as after all, even the widely popular neural networks were already conceptually developed in the 1960s).

XAI methods for transparent and post-hoc techniques

As stated, the dominant ML techniques currently used can be divided in two categories, either being transparent (interpretable) or requiring post-hoc analysis [Arrieta et al.]. The transparent techniques have the properties of simulatability, decomposability, and algorithmic transparency; they are:

- Linear/Logistic Regression
- Decision Trees
- K-Nearest Neighbors
- Rule Based Learners
- General Additive Models
- Bayesian Models

While the following techniques require post-hoc analysis (with their usually suitable method in parenthesis):

- Tree Ensembles (model simplification or feature relevance)
- Support Vector Machines (model simplification or local explanations)
- Multi–layer Neural Network (model simplification, feature relevance or visualization)
- Convolutional Neural Network (feature relevance or visualization techniques)
- Recurrent Neural Network (feature relevance)

Transparent models thus exhibit: "Simulatability: ability of a model of being simulated or thought about strictly by a human, i.e., transparency at the entire model level. Decomposability: every part of the model is understandable by a human without the need for additional tools, i.e., transparency at the individual component level such as parameters. Algorithmic transparent: ability of the user to understand the process followed by the model to produce any given output from its input data, i.e., transparency at the level of the training algorithm." [Arrieta et al.]

Post-hoc explainability can be realized as [Arrieta et al.; Lipton]:

- Text explanations: learning to generate text explanations that help explaining the results from the model.
- Visual explanations: visualizing the model's behavior.
- Local explanations: segmenting the solution space and giving explanations to less complex solution subspaces that are relevant for the whole model.
- Explanations by example: extraction of data examples that relate to the result generated by a certain model, enabling to get a better understanding of the model itself.
- Explanations by simplification: a whole new system is rebuilt based on the trained model to be explained.
- Feature relevance explanations: clarify the inner functioning of a model by computing a relevance score for its managed variables.

Other descriptions of post-hoc methods are [Guidotti et al.; Arya et al.]:

- Features Importance. E.g. the coefficients of linear models.
- Saliency Mask. Generally used to explain deep neural networks, can be considered a visual representation of FI. A visual representation of a layer of a neural network, e.g. to see what parts of an image had influence in the outcome.
- Sensitivity Analysis. It consists of evaluating the uncertainty in the outcome of a black box with respect to different sources of uncertainty in its inputs. It is generally used to develop visual tools for model inspection.
- Partial Dependence Plot. These plots help in visualizing and understanding the relationship between the outcome of a black box and the input in a reduced feature space.
- Prototype Selection. This explanator consists in returning, together with the outcome, an example very similar to the classified record, to make clear which criteria the prediction was returned. A prototype is an object that is representative of a set of similar instances and is part of the observed points, or it is an artifact summarizing a subset of them with similar characteristics.
- Activation Maximization. The inspection of neural networks and deep neural network can be carried out also by observing which are the fundamental neurons activated with respect to particular input records, i.e., to look for input patterns that maximize the activation of a certain neuron in a certain layer. AM can be viewed also as the generation of an input image that maximizes the output activation (also called adversarial generation).

For Deep Neural Network techniques several specific expl anation methods have been proposed: visualization methods, backpropagation-based methods, perturbation-based methods, model distillation, local approximation, intrinsic methods, attention mechanisms, and joint training [Xie et al.].

Importance of XAI in the AI design process
While various types of methods exist to facilitate XAI for various types of techniques, integration of XAI in the design process of the AI is a proposed way to offer fully explainable systems [Doran et al.]. Such an "intrinsic method" [Xie et al.] might even include the quality of explanation as an output variable to be optimized. Additionally, a modular approach can be used for XAI if it is considered early in the design process, to properly cater to the full range of stakeholders.

For Deep Neural Network techniques several specific explanation methods have been proposed: visualization methods, backpropagation-based methods, perturbation-based methods, model distillation, local approximation, intrinsic methods, attention mechanisms, and joint training [Xie et al.].

Importance of XAI in the AI design process

While various types of methods exist to facilitate XAI for various types of techniques, integration of XAI in the design process of the AI is a proposed way to offer fully explainable systems [Doran et al.]. Such an "intrinsic method" [Xie et al.] might even include the quality of explanation as an output variable to be optimized. Additionally, a modular approach can be used for XAI if it is considered early in the design process, to properly cater to the full range of stakeholders.

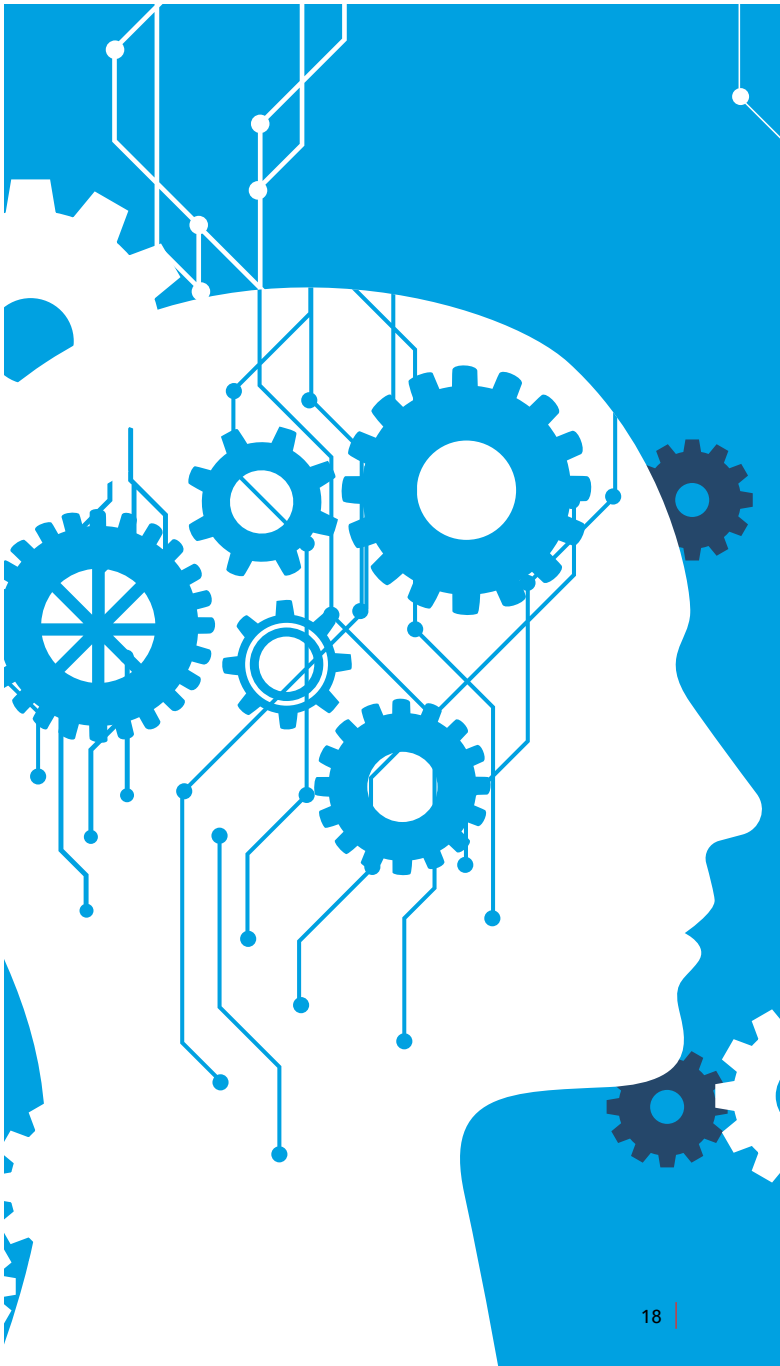
In the design process of XAI, the following aspects should be noted [ICO]:

- Select priority explanations (considering the domain, use case, and impact on the individual): know what type explanation you need prior to the design process of your AI solution.
- Collect and pre-process your data in an explanation-aware manner.
- Design the AI solution to be able to allow for various types of explanation.
- Translate the rationale of your system's results into useable and easily understandable reasons.
- Prepare implementers to deploy your AI solution. When human decision-makers are meaningfully involved in deploying an AI-assisted decision (i.e. a decision that is no fully automated), you should make sure you have appropriately trained and prepared them to use your model's results responsibly and fairly.
- Consider how to build and present your explanation. Considering how to construct clear and accessible explanations.

Evaluation of deployed XAI solutions

Evaluation of XAI methods can be desirable to see if it achieves the goal of offering understandable explanations. An evaluation process can raise the perceived trustworthiness of AI, if it is repeatedly demonstrated that XAI goals are met [Kozyrkov]. One author [Doshi-Velez et al.] described the various evaluation approaches:

- Application-grounded evaluation entails performing experiments using a functional AI application offering explanation. An expert (e.g. a medical doctor in the case of a AI that diagnoses patients) can evaluate whether the reasons given for a certain outcome line up with conventional heuristics.
- Human-grounded evaluation focuses more on a general user's scoring of the explanation provided by the AI, or their ability to indicate an output given input and explanation on the mechanics of the AI application.
- Functionally-grounded evaluation uses already validated AI techniques that underlie a novel application as a 'proxy' to establish it is also interpretable. While these evaluation methods are still somewhat abstract, they offer a potential framework to further develop evaluations for XAI. Similarly, to how valid XAI itself is context-dependent, the exact method of evaluating the quality of XAI might also be context-dependent.



8. LIMITATIONS OF XAI

Mittelstadt et al. warn that "explainable AI generates approximate simple models and calls them 'explanations', suggesting reliable knowledge of how a complex model functions" and relates this to Box's maxim: "All models are wrong, but some are useful". The wrong evaluations can lead to an XAI solution that seems explanatory, but actually misses crucial nuances. Therefore, appraisal of the given explanations should not be exclusively given by (one type of) the end-user.

Interestingly, while explainability can increase trust, explainability by itself is neither sufficient nor necessary for trust [Schaefer et al.; Sileno et al.]. A self-driving vehicle of which you perfectly know the algorithm and its operations, i.e. you have complete understanding of it at all times, will still be untrustworthy if it is unsafe. Conversely, many people trust things they do not fully understand the workings of, such as their smartphone or even their doctor performing an operation on them. In fact, it is not known how a human brain functions to reach decisions, yet we generally trust people's decisions. Nevertheless, explanations or reasons are often valued positively, but we should not stare blindly at a functionally complete, but incomprehensible explanation. Explaining exactly what circuits and logic gates were activated when your computer calculated something is technically informative, but practically not at the desired level of an average stakeholder. Trust does not always depend on understanding [Kozyrkov]. Section 5 covers why a proper explanation is selective.

Another limitation of XAI is that transparency can conflict with the broader goals of AI [Lipton], i.e. for some methods a trade-off might exist between transparency and performance. An XAI outcome can be 'gamed' so that it gives an explanation that pleases a stakeholder, e.g. a customer. On the other hand, an AI solution can be made deceptive, for instance for financial gain of a company [Mittelstadt et al.]. A certain transparency is called for that ensures no (unintentional) deception occurs [ICO; HLEG]. Finding the right balance between transparency and performance is not an easy task [Van der Burgt].

XAI could benefit from modularity, in which during the design process all modules be based not only on functional/mathematical techniques, but also on a set of principles or laws to justify its usage. This would also enable traceability and accountability in the final AI solution. Gartner also advises AI developers to foster ongoing conversations with all facets of their business, including the legal side [Sicular et al.]. Internal bodies that monitor AI (developers) legal and ethical compliance are good practice in any company using AI.

The Financial Stability Board (FSB) stated that the lack of interpretability of AI and ML methods are "a potential macrolevel risk" [FSB]. In addition, they warned that the widespread use of AI models that lack explainability could have unintended consequences [Sicular et al.]. Often it takes time before the general public trusts new technologies. For trains, electricity, and many other innovations the initial public reaction was caution. Over time, these technologies proved safe, as very few accidents happened and edge cases in which accidents happened were rare. However, with AI edge cases might have large effects, due to the potential reach of a certain AI solution. A flawed assumption in a widespread technology might lead to some problems. However, due to how pervasive AI is expected to be, the effects might be disastrous. Societies' conventional approach to deal with new technology's problems as they arise might not be suited for AI.

Currently we are (luckily perhaps based on the above section) in the 3rd AI 'hype cycle', some argue. Driven by new possibilities due to the increased speed of computers, results such as AI beating humans at Go, and impressive computer vision feats, we increasingly put trust in AI. However, we might have too much expectations of AI being the solution to all our problems. Similarly, we might expect too much of XAI as the solution of all our AI problems. Gartner states 5 Myths on Explainable AI: 1) Lack of explainability is a new problem specific to black-box AI; 2) All black-box AI must be explainable and interpretable; 3) Black-box AI decisions can be fully explained; 4) Human decisions are more explainable than black-box AI decisions; and 5) Explainable AI can be bought. AI should not replace human decision making, but rather augment and aid it [Alaybeyi et al.]. This also has a benefit as "Having humans make the ultimate decision avoids some complexity of explainable AI". The combined capacities of a human with an (X) AI system can be expected to outperform AI with no human help, especially due to the social nature of an explanation to a non-expert such as the end user.

9. A FRAMEWORK FOR XAI IN THE FINANCIAL SECTOR

While AI is expected to transform many facets of our lives in the coming decades, in the financial sector this might be especially pronounced in the coming years. The financial sector is consistently named as one of the sectors that invests most heavily in AI [DNB]. An extensive survey among financial institutions revealed that 77% anticipate that AI will be highly impactful [Ryll et al.]. According to the World Economic Forum (WEF), AI will transform the financial ecosystem by introducing new ways to distinguish financial institutions to the customers [McWaters]. The DNB-report gives the following examples of current AI use cases in the financial sector: "advanced chatbots, identity verification in client onboarding, transaction data analysis, fraud detection in claims management, pricing in bond trading, anti-money laundering monitoring, price differentiation in car insurance, automated analysis of legal documents, customer relation management, risk management, portfolio management, trading execution and investment operations" [DNB]. The Bank of England and the Financial Conduct Authority argue that ML is increasingly being used in UK financial services: "two thirds of respondents (of a survey) report they already use it in some form. The median firm uses live ML applications in two business areas. This is expected to more than double within the next three years. ML is most commonly used in anti-money laundering and fraud detection as well as in customer-facing applications (e.g. customer services and marketing). Some firms also use ML in areas such as credit risk management, trade pricing and execution, as well as general insurance pricing and underwriting" [BoE].

Some papers describe in more detail examples or approaches for XAI in the financial sector. Bracke et al. outline the different types of meaningful explanations one could expect from a ML model from a regulator perspective. A developer may be interested in individual predictions, for instance when they get customer queries but also to better understand outliers. Similarly, conduct regulators may occasionally be interested in individual predictions. For instance, if there were complaints about decisions made, there may be an interest in determining what factors drove that particular decision. Other stakeholders may be less interested in individual predictions. For instance, first line model checkers likely would seek a more general understanding of how the model works and what its key drivers are, across predictions. Similarly, second line model checkers, management and prudential regulators likely will tend to take a higher level view. Especially in cases where a model is of high importance for the business, these stakeholders will want good model development and governance practices across the board, the detail and stringency of standards on models vary by application. One area where standards around model due diligence are most thorough, is that of using models to calculate minimum capital requirements. Another example is governance requirements around trading and models for stress testing [Bracke et al.].

The WEF points to new ethical dilemmas that come along with the rise of AI. The WEF states that "the enigmatic nature of AI technology may seem like magic to outsiders but understanding its behavior is critical to detecting and preventing models that discriminate against or exclude marginalized groups and individuals" [McWaters]. In their report on the use of AI in the financial sector, De Nederlandsche Bank argues that "the use of AI in finance is special since the financial sector is commonly held to a higher societal standard than many other industries, as trust in financial institutions is considered essential for an effective financial system." [DNB]. For instance, responsible use of personal data, which are central in the use of AI, has been shown to be a major factor in trust in financial institutions [Van der Cruijsen et al.]. DNB introduced six principles of responsible use of AI. One of these principles is transparency which means that "financial firms should be able to explain how they use AI in their business processes, and (where reasonably appropriate) how these applications function" [DNB]. Lack of explainability is regarded as one of main risks of ML applications [BoE]. The ACPR regards explainability one of four evaluation principles for AI algorithms and models next to appropriate data management, performance, and stability [ACPR]. Explainability can be regarded as a means to enhance trust in the financial sector. Regulators attach great importance to this subject.

Arya et al. describe an automated AI lending scenario, identify stakeholders, and the types of explanations those stakeholders might require. A bank generally has an existing process combining a data-driven approach with business rules to arrive at loan approval decisions, which an AI solution (initially) will not fundamentally alter. Two stakeholders are easily identified: the loan officer, who wants to validate the approval or denial of loans given by the AI solution; and the applicant whose loan

was granted or denied. A third stakeholder is a data science senior or other internal party that would like to safeguard the lending process is up to standard. Arya et al. subsequently described which technical types of explanation best fits each party in her own taxonomy [Arya et al.].

Figure 4 contains a use case from the financial sector of how the conceptual framework from figure 3 can be applied. Inspired by Arya and colleagues we selected a use case of lending to consumers. The loan applicant in this use case is the end user. The financial adviser is the external adviser who gives the end user an explanation. The loan officer is the internal adviser who might provide the financial adviser with an explanation. We also included four different regulators in this example. The Netherlands Authority for Consumers and Markets (ACM) ensures fair competition between businesses and protects consumer interests. The Authority Financial Markets (AFM) is committed to promoting fair and transparent financial markets. As an independent market conduct authority, the AFM contributes to a sustainable financial system and prosperity in the Netherlands. The Autoriteit Persoonsgegevens (AP) is the Dutch Data Protection Authority who supervises processing of personal data in order to ensure compliance with laws that regulate the use of personal data. De Nederlandsche Bank (DNB) is the Dutch prudential supervisor who is committed to a reliable financial system, therefore it supervises banks, pension funds, insurers, and other financial institutions.

We put an "X" in certain cells of the framework to illustrate which type of stakeholder most likely requires which type of explanation in the case of lending to consumers. Please note that we talk about types of stakeholders. Particular loan applicants for example might require different explanations depending on their particular concerns. Some stakeholders, such as the domain expert or the AFM regulator, require a holistic view of the AI system, and thus require various pieces of information which in turn require various types of explanation. Other stakeholders are interested in a more narrow range of information, and require less (types of) explanation.

	Type of stakeholder												
	External		**Service provider**							**Other stakeholder (context)**			
Type of explanation	Loan applicant	Financial adviser	Loan Officer	AI Developer	Domain Expert	Executive mgmt.	Operational mgmt. (1st line)	Operational control (2nd line)	Audit (3rd line)	ACM	AFM	AP	DNB
The reasons, details of underlying causes of a particular outcome	X	X	X								X	X	
The data and features used as input to determine a particular outcome	X	X	X	X							X	X	
The data used to train and test the AI solution				X	X	X	X	X	X	X	X	X	X
The performance and accuracy of the AI solution				X	X	X					X		X
The principles, rules and guidelines used to design and develop the AI solution				X	X	X	X	X		X	X	X	X
The process that was used to design, develop, and test the AI solution					X		X	X	X				
The process of how feedback is processed					X		X	X	X				
The process of how explainers are trained					X		X						
The persons involved in design, development and implemantation of AI solution.					X	X	X	X	X				
The pesons accountable for development and use of AI solution						X	X	X	X	X	X	X	X

Figure 4. Conceptual XAI framework for lending to consumers.

10.CONCLUSION

In this paper we created a conceptual framework that can be used to map different types of explanations to different types of stakeholders for different use cases, with a focus on, but not limited to, the financial sector. The contribution of this paper is twofold. First, we provide practitioners with a framework that is useful in designing human-centric AI solutions. Second, with our definition and framework we add a broad and human-centric perspective on XAI to the existing body of research. Further research is required to validate the framework.

Foremost, we would argue that XAI is fundamentally a human-centric endeavor as ultimately it are humans that require explanations; rightfully so, given AI's potential impact in human life. An XAI solution should be designed with the receiving stakeholders in mind, as an explanation should not only be functionally complete, but also concur with the capacities of the stakeholder. When designing and building an AI solution our proposed framework can serve as a practical guide to determine what kind of explanations should be provided to what type of stakeholder. As a next step, the appropriate XAI method/technique can be selected depending on the AI system, taken into account all the other requirements for the AI solution (see the section on limitations). It should be noted that XAI is still a novel area of research and that much work remains to be done to smartly open up the black box of an AI solution, to enable proper explanations that address stakeholder concerns. Especially in domains like finance and healthcare, where complex and impactful decisions are made, a human in the loop is presumably required for the foreseeable future. Such a human explainer, who is capable of understanding and interpreting the outcome of the AI solution, can give a (layman) explainee a satisfactory explanation where a fully automated system can currently not. The potential benefit of a human explainer cooperating in the process underlines that XAI should be considered human-focused rather than a purely technical challenge. Ultimately, AI should be designed to assist mankind, not the other way around.

REFERENCES

- ACPR (2020). Governance of Artificial Intelligence in Finance. Retrieved from https://acpr.banque-france.fr/en/governance-artificial-intelligence-finance
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160.
- Alaybeyi, S., Linden, A. & Reynolds, M. (2019). 5 Myths about explainable AI. Gartner Research Note. Research ID G00464980.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
- Arya, V., Bellamy, R., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Vera Liao, Q., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K.R., Wei, D., & Yunfeng Zhang, D. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012.
- BBC (2014). Stephen Hawking warns artificial intelligence could end mankind. Retrieved from: https://www.bbc.com/news/technology-30290540#:~:text=Prof%20Stephen%20Hawking%2C%20one%20of,end%20of%20the%20human%20race.%22
- Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In IJCAI-17 workshop on explainable AI (XAI) (Vol. 8, No. 1).
- BoE (Bank of England) & FCA (Financial Conduct Authority) (2019). Machine Learning in UK Financial Services. Retrieved from https://www.bankofengland.co.uk/report/2019/machine-learning-in-uk-financial-services
- Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: an application to default risk analysis. Retrieved from https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis
- Cui, X., Lee, J. M., & Hsieh, J. (2019). An Integrative 3C evaluation framework for Explainable Artificial Intelligence.
- DNB (De Nederlandsche Bank) (2019). General principles for the use of Artificial Intelligence in the financial sector. Retrieved from https://www.dnb.nl/binaries/General%20principles%20for%20the%20use%20of%20Artificial%20Intelligence%20in%20the%20financial%20sector_tcm46-385055.pdf
- Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
- Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Vigneswara Ilavasaran, P., Janssen, M., Jones, P., Kumar Kar, A., Kizgin, H., Kronemann, B., Lal, B., & Williams, M.D. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 101994.
- FSB (Financial Stability Board) (2017). Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. Retrieved from https://www.fsb.org/wp-content/uploads/P011117.pdf
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.
- Gunning, D. (2017). Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web, 2.
- Hagras, H. (2018). Toward human-understandable, explainable AI. Computer, 51(9), 28–36.
- HLEG (The High-Level Expert Group on Artificial Intelligence) (2019). Ethics Guidelines for Trustworthy AI. EU Document. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
- ICO (Information Commissioner's Office) and Alan Turing Institute (2020), Explaining decisions made with AI. Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence/
- IIA (Institute of Internal Auditors) (2013). IIA position paper: the three lines of defense in effective risk management and control.
- Kozyrkov, C. (2018 'Explainable AI won't deliver. Here's why', Hacker Noon (16th Nov). Retrieved from https://hackernoon.com/explainable-ai-wont-deliver-here-s-why-6738f54216be
- Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31-57.

- McWaters, J. R. (2018). The New Physics of Financial Services: Understanding how artificial intelligence is transforming the financial ecosystem. In World Economic Forum.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.
- Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In Proceedings of the conference on fairness, accountability, and transparency (pp. 279-288).
- Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint arXiv:1902.01876.
- NIST (National Institute of Standards and Technology): Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020). Four Principles of Explainable Artificial Intelligence [Preprint]. Retrieved from https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence-draft
- NVB (2020). Artificiële intelligentie in de financiële sector. Retrieved from https://www.nvb.nl/media/3118/nvb-ai-in-de-financie-le-sector.pdf
- Ryll, L., Barton, M.E., Zhang, B., McWaters, J.R., Schizas, E., Hao, R., Bear, K., Preziuso, M., Seger, E., Wardrop, R. & Rau, P., Debata, P., Rowan, P., Adams, N., Gray, M. & Yerolemou, N. (2020). Transforming Paradigms: A Global AI in Financial Services Survey. SSRN Electronic Journal. DOI: 10.2139/ssrn.3532038.
- Schaefer, K., Chen, J., Szalma, J., & Hancock, P. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. Human Factors: The Journal of the Human Factors and Ergonomics Society, 58(3), 377–400.
- Sicular, S., Hare, J. & Brant, K. (2019). Hype cycle for artificial intelligence. Gartner Research Note. Research ID G00369840.
- Sileno, G., Boer, A., & van Engers, T. (2019). The Role of Normware in Trustworthy and Explainable AI. CEUR Workshop Proceedings, 2381, 9-16.
- Van der Burgt, J. (2020). Explainable AI in banking. Journal of Digital Banking, 4(4), 344-350.
- Van der Cruijsen, C., de Haan, J., & Roerink, R. (2020). Trust in financial institutions: A survey. DNB working paper. Retrieved from https://www.dnb.nl/en/binaries/Working%20paper%20No.%20693_tcm47-389881.pdf
- Xie, N., Ras, G., van Gerven, M., & Doran, D. (2020). Explainable Deep Learning: A Field Guide for the Uninitiated. arXiv preprint arXiv:2004.14545.

CONTACT

To give feedback on the whitepaper or to express your view on XAI in the financial sector you may contact us at martin.m.vandenberg@hu.n or ouren.kuiper@hu.nl.

You can also contact us in case you are interested in collaborating with us in XAI research. We are always open to sharing our research and to engage with interested organizations and researchers.

More information on our ongoing research on XAI in the financial sector can be found on our project page www.hu.nl/onderzoek/projecten/uitlegbare-ai-in-de-financiele-sector.