

Explicability Assessment Framework (EAF) for Automated Credit Decisions by Machine Learning Systems

N. J. Herber^{1, a}

¹ Faculty of Technology, Policy and Management, Delft University of Technology

Date: 24th of October 2019

Abstract

The use of machine learning systems has great potential to better predict probabilities of default for credit underwriting. Despite this advantage, herewith there exists the substantial risk of discrimination. Moreover, machine learning models with the highest prediction-accuracy are often the least explicable (i.e. explainable).

Nonetheless, *explicability* is needed to create accountability of automated credit decisions by machine learning systems. Furthermore, there exists a regulatory need for explicability of machine learning systems in the General Data Protection Regulation and the Consumer Credit Directive. Besides that, an ethical- and societal need exists for explicability.

Within the exploration of literature, it becomes clear that research lacks on how to move from a high-level principle like explicability, towards a prospective assessment of a machine learning use case on this principle, it lacks a multi-disciplinary perspective, and it misses an assessment framework that can guide decision-makers within machine learning use cases, aligned with a multi-organizational development lifecycle.

This research aims to design a prospective pragmatic assessment framework that can guide decision-makers, within machine learning applications in European credit underwriting cases from the point of view of explicability. To accomplish this, the Design Science Research Methodology, complemented with the Value Sensitive Design approach, is utilized.

To this end, the Explicability Assessment Framework (EAF) was developed. This framework is adapted to the context- and explanation characteristics of the case, and aligns with the CRISP-DM development lifecycle. It was found in two case studies that the framework helps with the decision-making whether a machine learning system is sufficiently explicable or not.

Lastly, a wide range of future research areas is identified that needs attention: empirical validation and expansion of the framework, the relevance for automated explanation creation, the scalability to other context and a large amount of explanations, and the practical perspective regarding adoption in the industry.

Keywords: *Explicability, Assessment, Machine Learning, Ethics, Credit Underwriting, Framework Design, Design Science Research Methodology, Value Sensitive Design*

1. Introduction

In the last couple of years research in the field of artificial intelligence (AI) has grown and more companies are acknowledging the advantages of this innovative technological area. The financial services industry sees it as

one of the most promising emerging technologies [1] and use-cases are already widely investigated for services, such as credit underwriting [2]–[5] and pricing of these services [6].

Machine Learning (ML), as a sub-category of AI, can be defined by combining the definitions of Russell & Norvig [7] and Samuel [8] as follows: *the use of self-learning algorithms*

^aSubmitted in partial fulfillment of the requirements for the degree of Master of Science in Complex Systems Engineering and Management. If additional information, substantiation or background is desired (such as the User's guidelines of the EAF, and the interviews), I refer to the Master's thesis that forms the base of this paper. This thesis can be found in the following repository: <https://repository.tudelft.nl/islandora/search/?collection=education>

from experience to adapt to new circumstances and to detect and extrapolate patterns. It has the potential to reduce costs, improve efficacy, find and create new business ventures and improve risk management [9]. In addition to these advantages, the application of this technique brings societal risks.

1.1. Problem background

The use of machine learning has shown that unpredicted and unwanted outcomes can occur. A problem that has occurred in several scenarios, is that human biases in observed data are being reproduced and even exacerbated by computers [10], [11]. In financial services, this could result in interest-rate discrimination [12] and a disproportional amount of expensive subprime loans issued to minority groups [13].

To deal with this problem, *accountability* is a first priority since this can help with doing justice when this occurs, and foremost, *accountability* should help to prevent these problems. In order to create *accountability*, an automated decision should be able to be explained [14]; i.e. a sufficient level of *explicability* is required.

Additionally, there is extra complexity with regards to the opaqueness of black-box models. The models with the best prediction-accuracy are often the least interpretable [15], [16], so there exists an important tension here: prediction-accuracy vs *explicability*.

The General Data Protection Regulation (GDPR) [17] as well as the Consumer Credit Directive [18] ask for the ability to explain decisions, with respectively the *right to explanation* and *right to non-discrimination*. To accomplish this, the machine learning system and the institutional system around it should be designed in such a way that the tension is systematically taken into account. Undesirable ethical, legal and societal effects should be minimized by designing the right kind and level of *explicability* in the systems; the *moral overload* [19] must be reduced. One way to do this is to improve the ML system based on a conducted evaluation of the ML system on *explicability* already in the development lifecycle. It is currently unclear how to structurally do this. Thus, the following research question guides the research:

“How can decision-makers prospectively assess machine learning applications within credit underwriting from the point of view of explicability?”

1.2. Recent progressions

Following the recent advancements and increasing interest of organizations in the ethics of AI, the European Commission appointed an AI expert group to advise (on a high-level) on the decomposition of the general strategy on Artificial Intelligence in Europe [20]. The advice from the group includes 4 principles towards achieving Trustworthy AI. The principles are Respect for human autonomy, Prevention of harm, Fairness, and *Explicability*.

This research focuses on the *explicability* of machine learning systems in credit underwriting for consumer loans (using machine learning models for risk prediction with a loan applicant). The lack of *explicability* of a system causes the problem that decisions cannot be justified to the subjects of these decisions, and therefore it remains unclear for them what the underlying reasoning is. This could clash with ethical principles (i.e. *accountability*, *transparency*, *human autonomy*) and even the law (i.e. GDPR, CCD).

In order to move from high-level principles (such as *accountability*, *explicability*) towards the pragmatic level of implementing these principles, a gap between these levels needs to be closed. A hands-on pragmatic assessment framework is developed by means of the Design Science Research Methodology (DSRM) [21] complemented with the Value Sensitive Design Approach [22] (these methods are explained in chapter 3). It takes a multi-disciplinary approach and is adapted towards the context characteristics, which is essential for AI ethics research [20], [23], [24].

1.3. Structure of the paper

Section 2 of this paper elaborates on the background of a few important aspects for the framework: machine learning in financial services, the values *explicability*, *transparency* & *accountability* for machine learning, the drivers for *explicability*, people related to ML *explicability*, a *good explanation* and the defined knowledge gap that this research fills. Section 3 describes the research approach that has been pursued. Section 4 elaborates on the results of the research. Section 5 discusses and evaluates the research and the results, after which section 6 concludes this research paper with answering the research question and describing the limitations and recommendations.

2. Background

2.1. Machine learning in the Financial Services Industry

Considering the amount of data that a bank owns, it is clear that utilizing this data for risk prediction is another area where using machine learning could create a lot of value. One of the interesting cases in this category for the foreseeable future is the application of machine learning models for credit underwriting. Banks see a lot of potential in using ML models for assessing the risk of a loan applicant in order to make more precise predictions of the probability of default, which cuts losses and increases the amount of approvals of good loans. The focus in this research is on personal consumer loans, due to the personal data that can be used and implicitly has the risk of discrimination.

This case has *accountability*, *trust* and *fairness* as the specific values of interest. *Trust* is lost in AI when users cannot understand the decisions of the systems [25]. A bank has an important *accountability* relation towards their

consumers and trust towards the banking industry is required for the stability of the financial system. Accountability is in this research the leading value, explicability as means towards this, but we keep in mind that in doing so we implicitly improve trust and fairness as well.

2.2. Explicability and machine learning

2.2.1. Explicability

Explainability and *explicability* are often used interchangeably in the scientific literature. The word explicability does not linguistically include the act of explaining, which makes it a more neutral word in comparison to explainability; it does not rely on a certain level of pre-knowledge, expertise on the subject or preferences of the explainee (the human who receives the explanation). This generality is required to ensure the explicability of a machine learning model that serves a wide range of knowledge among the explainees within this case, so therefore we choose explicability over explainability.

Within the scope of this paper, the definition of Gilpin et al. [26] complemented by the definition of Mittelstadt, Russell & Wachter [27] will be central: ‘*an explicable system is a system that can create complete and post-hoc human interpretable explanations of models and decisions, especially with respect to how it behaved, and why*’. In order to be able to explain a machine learning model, it is required that the information that concerns the model aspects of interest is available to investigate; it should be *transparent* enough for the explainer to inspect the workings and formulate an explanation based on this investigation.

2.2.2. Transparency

From the machine learning perspective, it is argued that transparency in itself shouldn’t be a goal, but a means to an end [28]. Explanations of actions (e.g. decisions) require transparency “*in terms of the algorithms and data used, their provenance and their dynamics, i.e. algorithms must be designed in ways that let us inspect their workings*” [14]. Thus, transparency can be seen as a means to explicability.

Transparency is the opposite of the opacity of black-box models [29]. Important to observe here is that models can be transparent without being explicable: a very complex model designed in such a way that all the internal workings can be inspected does not directly imply that its workings are understandable enough for an explainee. Transparency on its own is not a sufficient condition to achieve explicability.

The goal of transparency eventually determines to what extent a system needs to be transparent. This goal within this thesis is to improve explicability in order to improve accountability. Full transparency to the public is not required nor advisable here [28], [30]. In fact, full transparency to the

public could cause issues related to violations of privacy, undermined efficiency and property rights on algorithms [30].

2.2.3. Accountability

Accountability can be defined in the sense of a social relation [31]: *a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences*. To achieve this, accountability “*requires both the function of guiding action (by forming beliefs and making decisions) and the function of explanation (by placing decisions in a broader context and by classifying them along moral values)*” [14]. The function of explanation can be seen as the operationalization of explicability. Therefore, explicability is a means for accountability. Explanation is discussed as a tool towards ensuring accountability [32]. Accountability itself can be seen as a step towards fairness or non-discrimination, and this reaches even broader to values such as trust, security, safety, and privacy; accountability can be a powerful means towards reducing risks and harm [33].

We can project the characteristics of a public accountability relationship [31] on the credit underwriting case and outline the case according to these characteristics:

1. There is a relationship between the bank (actor) and the credit applicant (forum)
2. In which the bank is obliged
3. To explain and justify
4. His decision regarding the credit application (conduct)
5. The credit applicant can pose questions
6. Pass judgement (e.g. the decision seems based on data that is odd to play a role in it and could even imply discrimination)
7. And the bank may face consequences (e.g. legal consequences, or publicly released news-item that could cause damage to the brand or distrust)

The improvement of explicability simplifies committing to the third characteristic and is, therefore, an important step towards accountability.

2.2.4. The drivers for explicability

Additional to the aim to ensure accountability, the Consumer Credit Directive (CCD) and General Data Protection Regulation (GDPR) are the main drivers from the law perspective for requiring improvement of explicability. Moreover, there is an ethical and societal need for explicability.

The CCD [18] describes that “*the creditor must inform the applicant immediately and without charge of the results of such consultation and of particulars of the database*

consulted". Thus, the consumer has a right to know what the credit decision is based on.

The GDPR [17] states that there is a *right to explanation*, however, within the scientific community, there is a doubt whether the GDPR is legally binding, concerning the 'right to explanation' (article 22) [34]–[36]. We move beyond the limitations of this regulation since this research does not have the goal to just comply to the regulations, but to intrinsically improve the incorporation of important values at stake in the design process of machine learning systems. Moreover, the willingness of humans to understand decisions that concern their lives (human autonomy) is arguably a legitimate interest which should lead to an adjustment of GDPR (or an alternative regulation) such that it overcomes the challenges that GDPR currently faces concerning legally binding explanation requirements.

The ethical need concerns the values that the subjects of the decision care about: *personal autonomy, fairness, trust, accountability, transparency and explicability*. Non-commitment to these values could ultimately lead to social denial of using this technology. From the perspective of the company using the technology, it is important that the system complies to the values of the employees. Employees won't be willing to understand how the system works if they do not trust it, and subsequently they cannot explain the system.

The societal need concerns the institutions put into place to improve ethics. The banking code [37] let bankers say an oath that let them *put the interests of customers first and to maintain and promote trust in the financial sector*. The United Nations Human Rights [38] describe the importance of equality, fairness and non-discrimination in multiple articles. These give additional reasoning for the need for explicability.

2.2.5. People related to ML explicability

Explicability of machine learning systems is dependent on the stakeholders of these systems: different types of explanations may be needed for different actors in different contexts. In addition, there currently are methods which aim to improve the explicability of ML systems. To approach this from a more practical perspective, the widely used CRISP-DM lifecycle model (figure 1) [39], [40] will be expanded in the *evaluation phase* with the *explicability* perspective. This is the phase where the assessment framework takes its place.

This model is inter-organizational and in practice in the industry, and therefore of use with regards to our pragmatic goal for the industry of the framework.

In the evaluation phase of CRISP-DM, the decision is taken about what training- and ML model the best to be used is to continue to deployment. To do this, the model(s) should be evaluated on the performance of the model and on explicability as well. Four different groups are identified as stakeholders for this explicability.

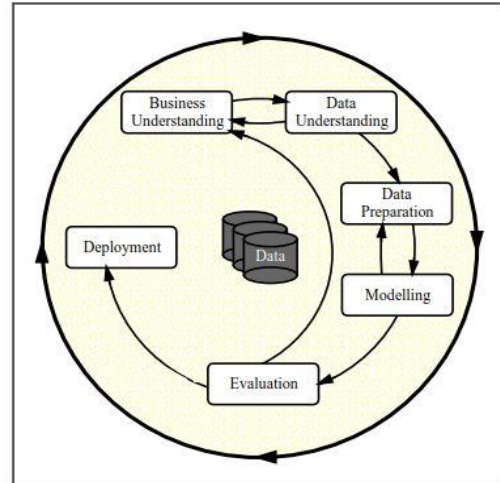


Figure 1: CRISP-DM

The model developers (data scientists) take on the evaluation task in the CRISP-DM, and is thus the target group of the framework (the users of the framework). The decision-maker within the evaluation step is the manager of the model development team.

On the other side of the spectrum, there is the layperson who receives an explanation on a credit decision that concerns him or her. The explanations for them need to be very accessible and understandable.

Further, there is the business employee within the company that use the ML system. They have a firm understanding of the business logic, reasoning behind the use and process, however, they miss the mathematical and technological knowledge; they require less mathematical explanations.

Lastly, there's the auditor who is an external controller and validator of the ML system after a system is deployed. They require explanations to check compliance with regulatory standards and principles.

2.3. What is a good explanation?

2.3.1. Explanation types

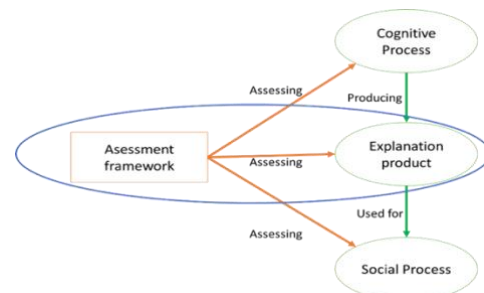


Figure 2: Assessing the 'explicability process'

There is a variety of explanation types that can be categorized in trifold [25]:

- Explanation as a cognitive process: “*the process of abductive inference to determine the causes of a given event, and a subset of these causes is selected as the explanation*” (formulating the explanation)
- Explanation as a product: “*the explanation that results from this cognitive process is the product*” (e.g. a textual, visual or conversational explanation)
- Explanation as a social process: “*the process of transferring knowledge between explainer and explainee (interaction) such that the explainee has enough information to understand the causes of the event*”

To fully assess explicability of a machine learning system, the development lifecycle has to be enhanced with tasks that assess all three categories of explicability, or as I call it ‘the explicability process’ (figure 2). However, for the scope of this thesis, the focus will be on the assessment of explanation as a product (circled with blue). The main reason for this is that it first needs to be clear what product the cognitive process requires to deliver, to understand how the cognitive process needs to be assessed, and how the social process needs to be designed and assessed.

The cognitive- and social process will in short be elaborated upon to set the context of the product, and to create a starting point for future research concerning the assessment of these two processes.

2.3.1.1. Cognitive Process

For future research the main question to be answered here is: *How can the methods used by the explainers to formulate a good explanation be assessed?*

The explaining methods can be classified among four categories, based on the problem that the explanation tries to solve with relation to a black-box model [41]:

- *The model explanation problem*: the formulated explanation by these methods is a human-interpretable and transparent model that mimics the behavior of the black-box model, e.g. (automatic) rule extraction [26]
- *The outcome explanation problem*: the formulated explanation by these methods is the specific rule that is used to classify a specific outcome of the model, e.g. LIME [42]
- *The model inspection problem*: the formulated explanation by these methods is a representation of some specific property of interest of the black-box model, e.g. SHAP [43], counterfactual explanation [41], [44]
- *The transparent box design problem*: the methods within this category are design methods for a transparent box model (opposed to a black-box model) and provide a human-interpretable and transparent model that can be used and does not require another model to mimic the behavior, in order to provide

locally or globally human-interpretable explanations, e.g. designing a decision tree classifier model [45]

This research focuses on the development of a framework that can be used for multiple types of cognitive processes (explaining methods).

2.3.1.2. Social process

Most literature does not incorporate the social aspects of explanation: transferring the (proposed understandable) formulated explanations to the explainee and the social interaction between the explainer and the explainee in relation to this. Nevertheless, alignment with societal values and ensuring the understanding of public opinion is very important [46].

The context of the case defines what social process needs to be put in place. The goal of the explanation is either *justification* or *teaching* [16]. Considering the accountability value of the credit underwriting case, the justification of a decision is the important goal for an explainee.

Further, validation is needed whether an explanation is understandable for an explainee. There are three different approaches for evaluation of explanations [32]: *application-grounded*, *human-grounded* and *functionally-grounded*. Considering the context of the credit underwriting case, the application-grounded evaluation seems the best fit, as there exists a concrete application. Moreover, social studies should investigate if there exists an optimal mix of these approaches for this case context.

2.3.1.3. Product

Explanation products can include different types of intelligibility [47]: *application*, *situation*, *input*, *output*, *model*, *why*, *why not*, *how*, *what if*, *what else*, *visualization*, *certainty*, and *control*. The context of the case for the assessment defines ultimately which of these types should be included [32]. In addition, this explanation product type leads to the explanation producing method (or multiple methods) that can be used.

Further, due to the justificatory goal, the explanation should answer the following question [48]: “*Why should we believe that this prediction is correct?*” Hence evidence on the correctness of the prediction (decision) should be included in the explanation. This can take the following forms and it is case-context dependent which ones to include [48]:

- *Normal* evidence and counter-evidence
- *Exceptional* evidence and counter-evidence
- *Contrarian* evidence and counter-evidence
- *Missing* evidence and counter-evidence

On top of that, an explanation product has a *moment in time* and a *scope* [34]: *ex-ante/ex-post* (prior to an automated decision and after an automated decision) and *global/local* (system functionality/specific decision)

Local ex-post explanation will be the scope within this research, as this type is the most relevant for the GDPR and CCD (justificatory value is needed required), for ensuring the explicability of this type is in line with the fourth characteristic of the public accountability relationship and since accountability is the upper value of interest.

Another dimension of explanation to be chosen is the *level of abstraction*. Three different levels are identified (figure 3).

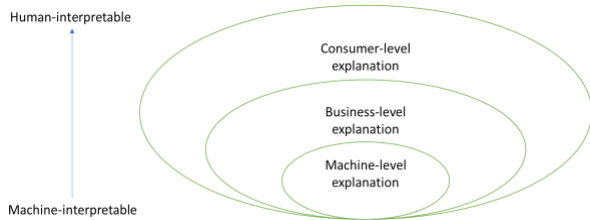


Figure 3: Levels of abstraction

- **Machine-level:** a mathematically sound explanation so that a data scientist and/or mathematician can validate the quality of the algorithms, calculations and model outcomes. On this level the explainers and the explainees are both humans from the data science domain who are required to have proficiency with regards to machine-interpretability.
- **Business-level:** an explanation that makes a decision understandable for the people in an organization whose tasks revolve around the value creation for the company and consumer. Within this level, the goal is to validate whether the model does what is supposed by the business objectives and if it can be implemented in the business processes. Within this level, the explainers are humans from the data science domain and the explainees are humans from the business domain of the organization.
- **Consumer-level:** an explanation that revolves around the data subjects of the decision, or the consumer, whose expertise level spectrum reaches from fully experienced to no experience with the content. This means that the explanations should be prepared to inform consumers from the full expertise level spectrum sufficiently. It focuses on human-interpretability and understandable language. Humans from the business domain are the explainers to the explainees in society (consumers).

The levels relate to each other and the confirmation that one level is validated forms the base for the other explanation, from bottom to top. A machine-level explanation must be of sufficient quality such that the mathematical and logical soundness can be validated. Without this, the decision-makers on the business-level do (legitimately) not trust the model enough to have it implemented. In addition, a high-quality business-level explanation is needed to be able to validate the soundness of this explanation in relation to the machine-level. This is required for the explainers to be able to ensure the

consumers that the used model excludes unwanted social inequities such as discrimination and non-fairness.

Concerning the GDPR, the consumer-level explanations are the most of interest and this is the scope of the research. The reason for this is the proposed (arguably non-legal) rights, in this regulation, for individuals that are the subject of the data that is used. The layperson perspective is the most interesting, since the consumer-level has the most to offer for them.

2.3.1.4. Explanation scope of interest

Table 1: Explanation assessment types

Class name	Characteristics of the explanation of interest	All possible characteristics of the class
Assessment moment	Prospective	Retrospective, prospective
Explicability process sub-part	Explanation product	Cognitive process, social process, explanation product
Explanatory value	Justification	Justification, teaching
Explanation scope	Local	Local, global
Moment in time	Ex-post	Ex-post, ex-ante
Level of abstraction	Consumer-level	Machine-level, business-level, consumer-level
Explainee	Layperson	Data-scientist, business, auditor, layperson

For the scope of this research, it boils down to the point that the assessment framework will be created for the characteristics stated in table 1. An important statement to make here is that not all characteristics can form logical combinations that are useful to assess: for example, the combination of local and ex-ante explanations, or machine-level explanations with a layperson as explainee

2.3.2. Explanation goodness evaluation

The quality of an explanation product can be evaluated among four identified characteristics [49], [50].

- **Completeness:** “the extent to which all of the underlying system is described by the explanation.” A very useful characteristic, but completeness does not imply directly more trust towards a system. This characteristic is more important for global explanations.
- **Conciseness:** “concise, easily consumable bites of information, that users can attend to if interested.” Always important, but has tension with completeness. An optimal point between those characteristics should be aimed for.
- **Soundness:** “the extent to which each component of an explanation’s content is truthful in describing the underlying system.” An essential characteristic and this should be aimed for within all types of explanations.

- **Comprehensibility:** selective in features to explain. An important aspiration, but an explanation can by default not be comprehensible for 100% of the society, so an important question here is to what extent this should be aimed for. Comprehensibility is extended with the linguistic aspects, such as choice of words.

The framework should make it possible to assess the explanation product on these four principles. The context properties, such as “the expertise of the user”, determine the level of importance of these characteristics [41]. If it turns out, within the execution of the development lifecycle, that a formulated explanation on a certain abstraction level is not good enough (i.e. the assessment results in the conclusion of a non-sufficient explanation), one should iterate back and try to find the cause, and improve the explanation. Moreover, if it turns out that the problem of inexplicability cannot be solved in a case where explicability is required, the data science team has to evaluate if another model that is possibly more explicable, can solve the problem and has a sufficient level of explicability.

2.4. Knowledge gap

The current literature shows that research on explicability in machine learning systems is emerging, however, some aspects are still underexposed and need research devoted to it.

Explicability has been shown to be a means towards more accountability of machine learning systems. Explicability is researched from a lot of different viewpoints (machine learning, philosophy, ethics, law), however it lacks research on how to prospectively assess explicability on a pragmatic level, which is needed in order to move from the ideals of incorporating values towards the real-world operationalization of values.

In addition, the literature on explicability often takes a mono-disciplinary viewpoint, and assessment requires a multi-disciplinary approach since the ‘goodness’ of explanation of a machine learning system has multiple perspectives. The contextual characteristics of the case ultimately define which aspects of explicability are required to be assessed, from which perspectives and this needs to be aggregated in the framework.

Further, the framework should be able to guide the design and development of a machine learning system with the decision-making concerning explicability of this system, so it needs to be aligned with the development lifecycle in place; literature lacks a robust framework that is aligned with a development lifecycle that is multi-organizational so that it can be used within multiple banking companies.

3. Research Methodology

As mentioned in chapter 1.2, this research uses the DSRM (Design Science Research Methodology) complemented by the VSD (Value Sensitive Design) approach. These methods

are chosen with regards to the goal of filling the knowledge gap.

3.1. Value Sensitive Design approach

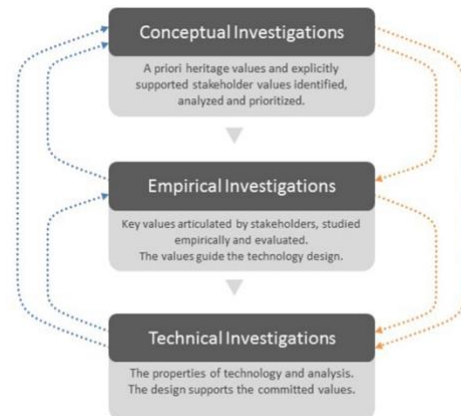


Figure 4: Value sensitive design tripartite [52]

Value Sensitive Design [22], or Design for Values, is an approach that enables the incorporation of human values throughout the whole design process. It builds on the theory that the impact of technology on the society is caused by its design features, the context of use and the users of the technology [51].

Within this research, VSD works in two ways:

First, the assessment framework will be designed for the values of accountability, with as means for this the value explicability, which eventually could enhance the values trust and fairness.

Second, the assessment framework should guide decision-makers within the development lifecycle in such a way that they incorporate the value of ‘explicability’ in their design of the machine learning system; the framework can be used to embed the value in the system.

Next to the designing for explicability and accountability, the assessment framework will be designed for using it in complex socio-technical systems; i.e. it will be designed for a bank. The use of machine learning the credit underwriting case requires the combination of technological, legal, ethical, social and business disciplines. Thus, an interdisciplinary approach is required to be able to address the aspects from different perspectives that come with explicability of machine learning systems in banks. The Value Sensitive Design approach is interdisciplinary by default and thus fits well to this need.

The focus is on the conceptual investigation from the VSD tripartite methodology (figure 4 [52]); the identification of the stakeholders, their values, and the guiding value for which the framework will be designed.

Since we focus on the design of a framework as the artifact, and not a purely technical artifact, the technical investigation is considered as less important. In addition, the empirical investigation of using the assessment framework is an aspect

for future research, although we include it a little by pursuing semi-structured interviews to validate the objectives. Real-world use of the framework should carry out a supported conclusion regarding the validity.

3.2. Design Science Research Methodology

The DSRM [21] provides a structured methodology for design research. Since the VSD-approach does not entail a functional methodology to design an artifact (the assessment framework, as deliverable) [53], the DSRM has been chosen as the main methodology, complemented by VSD. One of the outputs of the DSRM is an artifact, which can be an (assessment) framework [54].

Using this qualitative methodology has the advantages of literature-based, practical guidance and a model that provides in a presentation [21]. Moreover, it is a strong methodology for research in information systems and revolves around the proof that the artifact is actually useful in a specific case. The DSRM consists of a comprehensive process of six phases.

3.2.1. DSRM Phase 1: Problem identification and motivation

In this phase, the problem is identified and motivation is given of why this problem needs to be solved by means of the framework, and what the value of this is. The introduction of this paper has discussed this phase. The objectives of the framework should at least be formulated in such a way that this problem can be effectively solved by the framework.

3.2.2. DSRM Phase 2: Objectives of the solution

This phase identifies and discusses the objectives that the assessment framework needs to achieve. In addition, these objectives will be used for the evaluation phase (phase 5) to validate if the framework has solved the earlier stated problem. The research aimed to formulate assessment factors that are applicable for the specific case and explanation type of interest. Further, the objectives should ensure the quality of the framework. Conducted semi-structured interviews helped with the enhancement of the validity of the objectives. The literature overview from chapter 2 in combination with the formulated problem statement from chapter 1 form the input for this phase. The output of this phase is a list of objectives that the framework should achieve.

3.2.3. DSRM Phase 3: Design and development

First, within this phase, the objectives are transformed through specification, as described by van de Poel [55]. Furthermore, the contextual needs for the framework are transformed into design-requirements [56]. Lastly, semi-structured interviews improved the validity of these design requirements. The requirement for the design-requirements are specified by van de Poel (2013) as well: *the design-requirements should satisfy the upper norm or the objective*

The design requirements serve a '*for the sake of*'-relation with the upper-norm, or objective; this relationship will be leading the formulation of the design requirements.. The deliverable consists of a list of design-requirements for the framework.

Second, the design-requirements are creatively converted into an assessment framework as artifact. This is accomplished by creating a morphological chart [57] and synthesizing the means that are needed to commit to the design-requirements. As a result the Explicability Assessment Framework (EAF) was created. Furthermore, a user's guidelines was designed to support the use of the EAF.

3.2.4. DSRM Phase 4: Demonstration

In this phase, the EAF was applied to two cases in the same area of credit underwriting for personal loans. The first one is an explanation by rule extraction from a support vector machine algorithm [58]. The second one is a counterfactual explanation from a set of classification algorithms [44] (including black-box algorithms such as support vector machine with linear kernel (SVC) and multi-layer perceptron). With these case applications, the user's guidelines has been utilized to fill the framework.

3.2.5. DSRM Phase 5: Evaluation

The evaluation phase entails the evaluation of the designed artifact on the formulated objectives from phase 2.

First, the results of the demonstration are observed and compared with these objectives, after which a conclusion can be formulated to which extent the framework effectively meets the objectives.

Second, the demonstration process itself will be evaluated, since the objectives-based evaluation is partly dependent on this. Within this phase the limitations of the framework and the research are discussed as well. In the end, a decision will be made if the framework is good enough or that a back iteration to phase three has to take place to improve the artifact, before moving on to the communication phase, and recommendations for future research can be derived from these conclusions.

3.2.6. DSRM Phase 6: Communication

The last phase consists of the discussion on the research phases, results and process. Further, it includes guidelines for using the framework and the generalization possibilities will be discussed as well. The main research question is answered, conclusions, reflection on the research & these conclusions are formed and recommendations for future research are formulated. Moreover, a main part of this phase is the presentation of the results and the research in a scholarly or scientific format, which is the case by means of this paper.

4. Results

After thoroughly executing the methodology from chapter 3, this research produced multiple results. First, objectives are formulated that are aimed to solve the problem from chapter 1. Second, a design has been created with design-requirements and means to commit to these requirements (appendix 1). Third, the Explicability Assessment Framework is developed (appendix 3). Lastly, this framework is demonstrated by means of two case studies (appendix 4 and 5). Chapter 5 (evaluation) elaborates on the evaluation phase of the DSRM, including the limitations of the framework and the research.

4.1. Objectives of the solution

To formulate the objectives, it is first necessary to outline three important assumptions.

First, the assessment will solely focus on textual explanations.

Second, the assessment will solely focus on local ex-post explanations, and this implies that it concerns hypothetical explanations, since the assessment takes place before a real-world decision has been made (see figure 5).

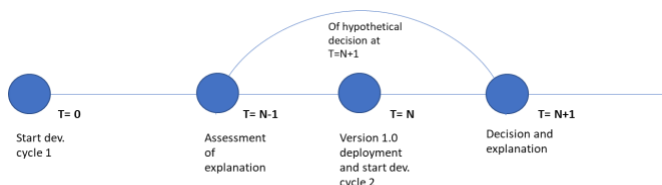


Figure 5: Timeline for the hypothetical decision

Third, decision-makers are assumed to choose a system that has better performance with a sufficient level of explicability over a system that has a worse performance with a sufficient level of explicability.

Next, the complexities are derived from the problem, the research objective and the literature overview. The design objectives are transformed from the complexities in such a way that the accomplishment of the objective(s) solves the complexities. The framework needs to reach the following 9 objectives:

- A. Provides guidance for the users of the framework
- B. Helps with decision-making on whether a certain task or decision-making functionality can be delegated to the machine learning system
- C. Is able to prospectively assess explanations
- D. Is able to assess justificatory explanations
- E. Is able to assess consumer-level explanations towards a layperson
- F. Is able to assess the completeness of explanations
- G. Is able to assess the soundness of explanations
- H. Is able to assess the comprehensibility of explanations
- I. Is able to assess the conciseness of explanations

Meeting these objectives result in a framework that solves the problem as formulated. The objectives are evaluated and validated within conducted semi-structured interviews with industry experts.

4.2. Design & Development phase

4.2.1. Design

Within this phase, design-requirements are developed to serve the higher the objectives [55]. The design requirements can be categorized as functional requirement (FR; *a service or function that the system should provide, a thing it should do, or some action it should take*), a non-functional requirement (NFR; *a quality, property or attribute that the system must possess*) or a constraint (C; a restriction or bound under which the system should operate, or the way in which the system is to be developed). A total of 20 design requirements are drafted, of which 18 are functional requirements and 2 are a constraint (see appendix 1). The requirements are validated by industry experts by means of semi-structured interviews.

4.2.2. Development

Subsequently, the transformation into a pragmatic prospective assessment framework continues with a morphological chart [56] with (a) means for every requirement (see appendix 1). The means are chosen such that for every requirement there is at least one means to meet this, and these are synthesized into the actual framework; the Explicability Assessment Framework (EAF).

4.2.2.1. Assessment framework relations

Before moving to this step, appendix 2 shows the relations and input that are of influence for the explicability assessment framework and shows more conceptually how we arrived at the point where we currently are in the development.

The context and explanation characteristics need to be documented for transparency of the situation and moreover define how the assessment framework needs to be adapted. This explicability assessment framework is adjusted to the *decision type* (the acceptance or denial of a credit application), the *(potential) effects* of this decision (receiving/being denied a loan, having more good outstanding loans/less bad outstanding loans, risk of discrimination) for *the actor* (the bank) and *the forum* (the loan applicant) of the *actor-forum relationship* (public accountability relationship). The *assessment moment* (prospective) is a choice that is led by pursuing the Value Sensitive Design approach and therefore the need to have a tool for assessment within the development lifecycle before development, such that the designers can make the right design choices. *The values* (accountability and explicability as means for accountability) lead, in addition to this assessment moment, to the *ethical need* in society (right

for explanation), *the regulations* (Consumer Credit Directive, GDPR) and the actor-forum relationship.

Next, the context characteristics interrelate with the explanation characteristics. The actor-forum relationship defines who the *explainee* (layperson) is of the explanation that is being assessed. Further, this explainee defines the *level of abstraction* (consumer-level) of the explanation and the *explanatory value* (justification) that is the most valuable for this explainee. The explanatory value defines the *explanation scope* (local) and the *moment in time* (ex-post) of the explanation. Lastly, the *explicability process sub-part* (explanation product) defines the assessment object, and is chosen.

Moving to the final part, the input for the Explicability Assessment Framework (EAF) is the *assessment object* (the explanation product), which in this case is a textual explanation. The framework needs to be aligned with and adopted with the *CRISP-DM lifecycle*. In order to do this, the framework is supported by the *user's guidelines*. The *Explicability Assessment Framework* results in an overview of the assessment and a final *conclusion* concerning the explicability of the machine learning system regarding the specific explanation and context.

4.3. Demonstration phase

The designed framework (appendix 3) is applied to two use cases supported by the user's guidelines.

Within the first case study, it becomes apparent that the explanation does not suffice. To improve explicability of the machine learning system the explanation should at least improve on *completeness, comprehensibility and the layperson perspective*. To do this, first the cause of contrarian evidence must be found and investigated in order to find out if it is a problem, and further the explanation should be made simpler, in a narrative format and should include a why-statement and more contextual details.

The second case study shows that the explanation does not suffice as well. To improve this, the contrarian evidence cause should be investigated to check if this is a problem, the explanation should be made simpler, in a narrative format, and include more of the logical links between the separate parts of the explanation, it should include a how-statement and context needs to be added.

Concluding, the application of the framework shows hands-on next steps for the decision-makers in the development lifecycle to improve explicability of the machine learning system, so it is pragmatic to use in combination with the user's guidelines. Further iterative steps should discover if these improvements can be made, and if thus the machine learning system can be made explicable within the specific context.

5. Evaluation

This chapter elaborates on the demonstration phase of the DSRM; the evaluation phase. This is an essential part since it focuses on serving the real-world objectives as defined.

First the EAF is evaluated on the achievement of the objectives. The question to answer here is: 'does the demonstration show that the framework accomplished the objectives?'.
Second, a broader scope is taken and the limitations of the framework and the research are discussed.

5.1. Evaluation of the EAF on the objectives

A. Provides guidance for the users of the framework

Yes, there is extensive guidance for decision-makers to use the framework, however, the complete usefulness still needs to be empirically validated in the industry by the proposed users of the framework.

First, the guidelines describe how the use of the framework fits within the CRISP-DM lifecycle. Although this model is often (slightly) adjusted to the specific use-case context, we argue that since this lifecycle model is a widely adopted one in the industry, many development lifecycles include the different phases of the lifecycle in some way; an evaluation phase where the assessment of explicability should take place. The guidelines, therefore, show the decision-makers when the assessment should take place.

Second, the guidelines include a step by step task description of what to fill in what section of the framework, including examples and what choices to make. In addition, it shows the sequence of the tasks to do. Finally, the guidelines include a recommendation of what to do as the next step with an insufficient level of explicability of the machine learning system (iterate back to the business understanding step).

B. Helps with decision-making on whether a certain task or decision-making functionality can be delegated to the machine learning system

Yes, the concluding section that summarizes the outcome of the assessment with the framework, complemented with the future steps to take, can ultimately form a decision if the task can be delegated to the machine learning system.

Both of the case studies show that a substantiated conclusion can be formulated based on the before conducted assessment. An enumeration is given of the different assessment subsections that are insufficiently good that lead to the conclusion that the explanation is insufficiently good. Moreover, it shows what parts need to be improved in order to improve the explicability, and this helps the decision-makers with the future steps to take.

C. Is able to prospectively assess explanations

Yes, the demonstration shows that it is possible to assess explanations for hypothetical decisions with the framework.

The demonstration uses two papers that both focus on machine learning systems that are applied to historical data in order to investigate the systems. The generated explanations are based on this as well, so since these systems are not deployed yet and hypothetically make the decisions, we can state that the explanations are on hypothetical decisions. The assessment of these explanations was successful, thus we conclude that the framework is able to be used to prospectively assess explanations. A prerequisite for this is that hypothetical decisions can be produced in order to be able to create local ex-post explanations.

D. Is able to assess justificatory explanations

Yes, with the selection of the right evidence roles and intelligibility types for the justificatory goal, the framework can be adjusted towards this goal, and thus reach this objective. Future empirical research should investigate among explainees which options are the most relevant.

The demonstration shows that for both cases the choices for evidence roles and intelligibility types can be made and supported and that this impacts question 1.X and 3.X in the assessment since these questions should be answered for the chosen options. The finalized frameworks in the demonstration show that the assessment of the choices for evidence roles and intelligibility types are essential in ensuring a good explanation that can be used for justification. The framework is, therefore, able to be used for assessment of justificatory explanations.

E. Is able to assess consumer-level explanations towards a layperson

Yes, the framework shows in the demonstration that it can be adjusted towards this perspective with the following parts: the selection of evidence roles and intelligibility types, question 2.1 and 2.2 in the framework and the comprehensibility and conciseness sections. Only the basics of the evidence roles and intelligibility types are therefore chosen in the demonstrations. In addition, the knowledge of a layperson is estimated with regards to the given explanation and an assessment takes place if this corresponds with each other. Concluding, we can say that the framework is able to assess consumer-level explanations towards a layperson.

However, since it is hard to say for another person what knowledge a layperson has and has not (due to a lack of information), empirical research should be conducted in order to investigate this, and a feedback-loop should be created with this new information to improve the explanations.

F. Is able to assess the completeness of explanations

Yes, the framework can be used to check which intelligibility types are included in the explanation.

The demonstration shows that first, the intelligibility types [47] have to be chosen that need to be included in an

explanation to be considered complete. Afterwards, the assessment can take place where we look if the intelligibility types are present in the explanation. Both cases show that the completeness of the explanation lacks and that additional information should be added in order to have a sufficient level of completeness of the explanation. We can conclude here that the framework has the ability to be used for the assessment of completeness of an explanation.

A note has to be made here that, even though the choices have to be supported, the chosen intelligibility types are still chosen, and the quality of this aspect can be improved when a certain standard is adopted so that it is directly clear in what situations what intelligibility types need to be included. Future research should investigate this aspect.

G. Is able to assess the soundness of explanations

The framework is able to partly assess the soundness of the explanations, with regards to the known workings of the system.

However, it is therefore required that the user of the EAF knows the workings, or that the user of the EAF can trust on the conclusion of someone else that can validate the inner workings. Thus, the machine-level validity and business-level validity of explanations on these levels need to be ensured, before we are able to ensure that the explanation is fully sound on the required level. The assessment of explanations on this level is out-of-scope in this research and requires research in order to design a framework to accomplish this (or extend this framework with this point of view).

Within the demonstration phase, we can see that section 4 of the EAF contains four different levels: *full truth model*, *simplified truth model*, *the truth of (a) singular feature(s) and not the truth*. It is evident that all the different subsections of an explanation need to be based on the truth; false, incorrect or untruthful statements cannot be tolerated. However, there can be situations where the explainees are not interested in knowing all the steps, all features and the full truthful workings of the system, but just parts of it. Both explanations in the cases possess solely statements that are truthful; the first explanation concerns a simplified truth model that mimics the full truth model, and the second explanation concerns the truth of a few singular features that the decision is based on.

H. Is able to assess the comprehensibility of explanations

The framework includes important aspects for assessment of the comprehensibility, however, future research is needed to fully understand the needs of laypersons regarding comprehensibility, since these are currently assumptions.

Within the demonstration, we can see that the assessment of these aspects in both of the explanations results in recommendations for improvement. Besides that, it is important to have future research conduct empirical surveys

among the explainees, with as final goal to design a feedback loop for improvement of explanations. This falls under the social process of explanation which is out-of-scope for this thesis.

I. Is able to assess the conciseness of explanations

The framework is able to document and assess the conciseness of explanations, however it is currently still unclear what the threshold is with this aspect of explanation. This should be researched in order to be able to fully assess the conciseness.

The framework contains three metrics to assess the length of an explanation: textual lines, amount of words and number of concepts. In addition, the aspect of modularity of an explanation plays a role here, since easily extendable explanation can first be very concise and if the explainee needs more explanation this can be added (this is an important aspect of the social process of explanation).

The demonstration shows that the framework is able to be used to acknowledge and document the facts concerning the length of an explanation and the modularity of an explanation. We, therefore, can conclude that the framework can be used for conciseness assessment, with the prerequisite that the users of the framework have a clear view on what the threshold is for a sufficiently concise explanation.

Moreover, future research should be focused on what the ideal situation is regarding the length of an explanation in the specific context, or what best-case ranges are. This is currently lacking and therefore no conclusion can be given regarding whether the explanation is concise enough. Empirical evidence is very useful for this aspect

5.2. Limitations

In this research, and especially after the conducted evaluation with regards to the objectives, it becomes clear that the framework has a lot of advantages that positively contributes to the design field of machine learning systems. However, the performed research comes as well with its limitations. There have been five limitations identified.

First, a limitation resulted from the ambitious aim to cover and synthesize a wide range of literature in the novel research area of explicability for machine learning into one framework. There was not a framework in the literature present what this framework could build upon, so a new framework had to be designed from scratch. Due to time and resource constraints, the framework as the output of this research covers just a modest part of the full explicability spectrum to cover (which is needed to fully ensure explicability of a machine learning system). It is a firm step in the right direction, but there are several prerequisites for the use of the framework, such as that the soundness of the explanation can only be ensured if the user of the framework is confident that the machine-level and business-level explanations are sound, and it has very specific

context characteristics. But then again, this framework is able to be used to assess the soundness with regards to what is known about the other levels. So, given that the other levels are *sound* (which should be assessed as well), the framework can be used to assess the soundness of an explanation.

The second limitation that has been identified is that the *evidence roles* and *intelligibility types* are currently chosen. Although the choices can be supported with arguments, there is a risk that a choice is hard to substantiate. Subsequently, there is a risk of confirmation bias, in such a way that users choose their evidence roles and intelligibility types strategically so that an explanation seems better than it actually is, and the machine learning system passes this specific assessment.

Additional to this second limitation, is the following limitation that unfolds from the evaluated demonstration: some factors of the framework (*soundness, conciseness*) lack a robust threshold to be used to observe if an explanation is good enough on this factor. Subsequently, this means that with the current framework the user should think of this by itself and this implies the risk of confirmation bias of the created explanations. Nonetheless, one can state that supervisory organizations should develop standards for this that can be implemented afterwards. The sections are present and should ultimately be extended with a minimum threshold to test the explanation on.

The fourth limitation is that the usability of the framework in combination with the user's guidelines is *not empirically validated by industry experts*, solely the design objectives and design requirements. Although the designer of the framework is able to apply the framework in the demonstration phase to two cases, it cannot directly be deduced that this is easily doable for users that did not design the EAF. Considerable effort has been taken to make it as pragmatic as possible, but this needs to be validated in future research. Current resource and time constraints do not allow to include this in the current research project.

Lastly, a limitation is the fact that it is hard for the decision-makers of a machine learning system to know what can be considered as the knowledge level of a '*layperson*'. First, because a layperson is actually a large group of diverse people, and secondly because without interaction with this group it is mainly a guess without validation what this knowledge is. Subsequently, it is hard to fully assess the comprehensibility towards a layperson of an explanation without including this knowledge from the layperson.

6. Conclusion

This final chapter presents the conclusions of the conducted research for this thesis. All sequential steps of the DSRM are executed and the results are synthesized and discussed here.

6.1. Main findings

By synthesizing and analyzing all the results from the prior steps, we can now answer the main research question:

“How can decision-makers prospectively assess machine learning applications within credit underwriting from the point of view of explicability?”

The demonstration and evaluation phases show us that the framework positively contributes to the ability of the machine learning system designers and decision-makers to evaluate the explicability of the system.

First, they need to describe and document the context characteristics of the use case, to create transparency and overview of the influential factors.

Second, the explanation characteristics need to be chosen. Concerning the applicability of the EAF, the explanation to be assessed should fall into the following scope: a textual explanation product (explicability process sub-part), ex-post (moment in time), local (explanation scope), justification (explanatory value), consumer-level (level of abstraction) and layperson (explaineé).

Next, framework adjustments have to be made to tailor the EAF to the context. Evidence roles and intelligibility types are chosen to commit to the justificatory value, consumer-level and layperson perspective.

Subsequently, the actual assessment can take place where the explanation is systematically assessed with questions on the justificatory value, the layperson perspective, completeness, soundness, comprehensibility, and conciseness.

The conclusion that is derived from this assessment can now be supported with arguments why a certain explanation is not good enough, and therefore why the machine learning system is not explicable enough.

Furthermore, these arguments create a starting-point of an improvement iteration towards a more explicable machine learning system.

6.2. Interpretation of the main findings

The answer to the main research question indicates that the findings contribute to a more systematical and theory-based assessment of explicability of machine learning systems in credit underwriting. This additional evaluation is needed to ensure a future-proof method for explicability compliance, with regards to the GDPR, CCD and foremost to satisfy the ethical need in society. The design of this framework is a significant step towards the ability to fully assess the explicability of a machine learning system.

Reducing the gap between explicability (a high-level principle) of the AI ethics literature and the operational-level of machine learning system development for credit underwriting is a novel field within the literature and has not been extensively addressed before. By means of an assessment

framework, a design-perspective has been taken, such that this novel evaluation method can be incorporated in the (CRISP-DM) development lifecycle: a prospective assessment tool that guides decision-makers and eventually helps with the improvement of explicability of machine learning systems.

In addition, this thesis contributes to two main action points from the *Ethical Framework for a Good AI Society* [46], the second and the fourth one:

- 2. *“Assess which tasks and decision-making functionalities should not be delegated to AI systems, through the use of participatory mechanisms to ensure alignment with societal values and understanding of public opinion. This assessment should take into account existing legislation and be supported by an ongoing dialogue between all stakeholders (including government, industry, and civil society) to debate how AI will impact society opinion.”*

- 4. *“Develop a framework to enhance the explicability of AI systems that make socially significant decisions. Central to this framework is the ability for individuals to obtain a factual, direct, and clear explanation of the decision-making process, especially in the event of unwanted consequences. This is likely to require the development of frameworks specific to different industries, and professional associations should be involved in this process, alongside experts in science, business, law, and ethics.”*

When we look at the generalizability of the findings, we can observe a few things. First, the hypothesis is that, despite the narrow scope of application, with just small adjustments of the framework, it can be possible to extend the scope and cover a wider range of explicability. Second, it might be very interesting for other application areas, that have similarities regarding the context and explanation characteristics, to investigate if this framework could help them as well with the evaluation of explicability in their ML systems.

However, the application scope of this framework is just a small part of the full explicability range, it should be seen as a step in the right direction, and this thesis can be qualified as a starting point for future research, in order to ultimately be able to cover the full range.

6.3. Recommendations for future research

The evaluation phase made clear that there are six main limitations with the conducted research. This paragraph contains recommendations for future research to cope with these limitations (in the same order as these limitations in chapter 5.2). Subsequently, it adds more general recommendations for future research, that relates to the generalizability and broader application field that is interesting to investigate in order to create value for the scientific community and society.

6.3.1. Recommendations regarding the limitations

To cope with the scope limitation of the framework, future research should investigate the change of an explanation characteristic one-by-one. The two main questions here are: *what does the change of an explanation characteristic influence regarding the explicability assessment framework*, and *how can the framework be adapted such that the new situation can be assessed?* When this is investigated, the researcher can adapt and design another framework that complements the EAF. By continuing this, a full set of frameworks can theoretically be designed (or one large framework) that should cover the full explicability scope.

Second, additional research must be conducted on what *intelligibility types* and what *evidence roles* to include in the explicability assessment with certain characteristics, to reduce the risk of confirmation bias. Currently, this is chosen, and the framework improves in robustness if there is a standard for this.

Third, the soundness and conciseness factors of an explanation require a certain threshold, so that the users of the framework know whether an explanation is good enough on those factors. However, currently, research has to take place on what the thresholds are in what context, such that the EAF can be adapted towards this.

Fourth, the usability of the framework should be validated by the actual proposed users of the framework. The question to be answered here in empirical research is: *Is the EAF easy enough to be used in real-world cases?* Despite the fact that the researcher can easily apply the framework, this does not directly imply a high usability level for all the users of the framework and this should be empirically tested.

Fifth, research should be conducted to empirically investigate what can be considered as the knowledge level of a layperson within this field. This should increase the ability to assess the comprehensibility of an explanation, and this increases the validity of the framework. The approach of Doshi-Velez & Kim [32] can be taken as starting point and a mix of *application-grounded*, *human-grounded* and *functionally-grounded* evaluation can be taken to investigate this aspect.

6.3.2. Further recommendations

Additional to the former recommendations regarding the limitations, there are several other research fields related that are interesting for future researchers as well.

To start off, it would be recommended to investigate the possibilities of this framework in another context, such as automated court decisions. Could this framework be adapted towards this context such that explanations on those decisions can be assessed? The ethical load and challenges are different, but one can see that there are similarities that might indicate the usefulness of this framework for this context.

Second, the scalability of the assessment is a very interesting field and research on how to assess not just one but

a whole range of explanations in as less time as possible would significantly improve the value of this research. Most people do not want to assess just one specific explanation every time, but multiple explanations first, and maybe certain outliers afterwards.

Further, it is advised that future researchers investigate what this systematical and structured assessment means for the automation of explanation generation (e.g. for this specific case). Could it be possible that next to the automated decision, a machine learning system is supplemented by an IT system that correctly generates good explanations that satisfy explainees, and ensure the explicability of the system? The outcome of this thesis is a good starting point for such a question.

To create even more robustness and validity for this research, a second round of expert interviews should be executed, in order to validate the framework. In addition, it would be interesting to explore more use cases and evaluate how the assessment goes with these cases.

Another research field that is advised to be investigated, is confirmation bias within the assessment itself, as this is executed in the evaluation phase of the development lifecycle by the designers of the ML system. Research has to be conducted in the field of explicability incorporation in the development lifecycle and the relation between confirmation bias and the quality of the assessment.

Lastly, the conducted interviews have shown us two interesting observations:

First, there exists a gap between the practical status quo of machine learning model development and the proposed assessment. The main added value of this research was searched in the quantitative field by the interviewees, however, this is a qualitative thesis and qualitative assessment should already be incorporated in the model development phase, which is aligned with the Value Sensitive Design approach. The need to incorporate this qualitative part at the beginning of the development was not directly seen by the interviewees. Future research should find out if this is a common thing in the industry, and how to create a change of perspective of these industry experts such that this need is seen by them as well.

Second, it appears that large banking companies are currently quite risk-averse with using machine learning systems in the credit underwriting cases. It obviously is a cost-benefit tension for the companies and it should be investigated what the costs are for fully 'implementing' explicability in such a machine learning system, and if it is still profitable for the banks to use machine learning systems in these cases, concerning the compliance to regulation and ethical need.

Acknowledgements

This paper has been written as part of my master's thesis for the MSc Complex Systems Engineering and Management at the Delft University of Technology. I had supervision from a graduation committee that guided me along my writing process:

- Chairperson: Prof. Dr. M.J. van den Hoven
- First Supervisor: Dr. F. Santoni de Sio
- Second Supervisor: Dr. S. Cunningham
- Advisor: PhD candidate S. Robbins, MSc.

Additionally, I had guiding supervision from J. Schijven, MSc., who is a manager at EY. Further, the thesis has been written as part of an internship at EY. They provided me with funding and access to a broad network in the financial services industry. EY had no control over the content of the thesis nor this paper.

References

- [1] R. R. Trippi and E. Turban, Eds., *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance*. New York, NY, USA: McGraw-Hill, Inc., 1992.
- [2] H. Ince and B. Aktan, "A Comparison of Data Mining Techniques for Credit Scoring in Banking: A Managerial Perspective," *J. Bus. Econ. Manag.*, vol. 10, no. 3, pp. 233–240, Sep. 2009.
- [3] A. Lui and G. W. Lamb, "Artificial intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector," *Inf. Commun. Technol. Law*, vol. 27, no. 3, pp. 267–283, Sep. 2018.
- [4] J. Boillet, "Why AI is both a risk and a way to manage risk," 01-Apr-2018. [Online]. Available: https://www.ey.com/en_gl/assurance/why-ai-is-both-a-risk-and-a-way-to-manage-risk. [Accessed: 13-Mar-2019].
- [5] ZestFinance, "ZAML® Credit and Risk Modeling Solutions | ZestFinance." [Online]. Available: <https://www.zestfinance.com/zaml>. [Accessed: 05-Mar-2019].
- [6] R. Balasubramanian, A. Libarikian, and D. McElhaney, "Insurance 2030--The impact of AI on the future of insurance | McKinsey," Apr-2018. [Online]. Available: <https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance>. [Accessed: 13-Mar-2019].
- [7] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, New Jersey, NJ, USA: Pearson Education, Inc., 2010.
- [8] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Dev.*, vol. 44, no. 1.2, pp. 206–226, Jan. 2000.
- [9] M. Berthold and D. J. Hand, *Intelligent Data Analysis: An Introduction*, 2nd ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.
- [10] J. Angwin, M. Varner, and A. Tobin, "Facebook Enabled Advertisers to Reach 'Jew Haters,'" *ProPublica*, 14-Sep-2017. [Online]. Available: <https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters>. [Accessed: 13-Mar-2019].
- [11] K. Crawford, *The Trouble with Bias - NIPS 2017 Keynote*. Long Beach, California, United States of America, 2017.
- [12] R. P. Bartlett, A. Morse, R. Stanton, and N. Wallace, "Consumer Lending Discrimination in the FinTech Era," *SSRN Electron. J.*, 2017.
- [13] C. Havard, "'On the Take': The Black Box of Credit Scoring and Mortgage Discrimination," *Boston Univ. Public Interest Law J.*, vol. 20, no. 2, pp. 241–287, Spring 2011.
- [14] V. Dignum, "Responsible Autonomy," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia, 2017, pp. 4698–4704.
- [15] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, 5th ed. New York: Springer Science & Business Media, 2013.
- [16] R. Sheh and I. Monteath, "Defining Explainable AI for Requirements Analysis," *KI - Künstl. Intell.*, vol. 32, no. 4, pp. 261–266, Nov. 2018.
- [17] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*, vol. 119. 2016.
- [18] *Directive 2008/48/EC of the European Parliament and of the Council of 23 April 2008 on credit agreements for consumers and repealing Council Directive 87/102/EEC*. 2008.
- [19] J. Van den Hoven, G.-J. Lokhorst, and I. Van de Poel, "Engineering and the Problem of Moral Overload," *Sci. Eng. Ethics*, vol. 18, no. 1, pp. 143–155, Mar. 2012.
- [20] European Commission, "High-Level Expert Group on Artificial Intelligence," *Digital Single Market - European Commission*, 14-Jun-2018. [Online]. Available: <https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>. [Accessed: 19-Apr-2019].
- [21] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *J. Manag. Inf. Syst.*, pp. 45–77, 2008.
- [22] B. Friedman, P. H. Kahn, A. Borning, and A. Hultgren, "Value Sensitive Design and Information Systems," in *Early engagement and new technologies: Opening up the laboratory*, vol. 16, N. Doorn, D. Schuurbiens, I. van de Poel, and M. E. Gorman, Eds. Dordrecht: Springer Netherlands, 2013, pp. 55–95.
- [23] M. Hansson, "Artificial Intelligence for Humans," *Think*, vol. 1, p. 20, 2019.

- [24] A. Winfield, "Ethical standards in robotics and AI," *Nat. Electron.*, vol. 2, no. 2, pp. 46–48, Feb. 2019.
- [25] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.
- [26] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018, pp. 80–89.
- [27] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining Explanations in AI," *Proc. Conf. Fairness Account. Transpar. - FAT 19*, pp. 279–288, 2019.
- [28] A. Weller, "Transparency: Motivations and Challenges," in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Muller, Eds. Cham: Springer International Publishing, 2019, pp. 23–40.
- [29] Z. C. Lipton, "The Mythos of Model Interpretability," *Queue*, vol. 16, no. 3, pp. 30:31–30:57, Jun. 2018.
- [30] P. B. de Laat, "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?," *Philos. Technol.*, vol. 31, no. 4, pp. 525–541, Dec. 2018.
- [31] M. Bovens, "Analysing and Assessing Accountability: A Conceptual Framework," *Eur. Law J.*, vol. 13, no. 4, pp. 447–468, 2007.
- [32] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *ArXiv170208608 Cs Stat*, Feb. 2017.
- [33] H. Nissenbaum, "Accountability in a computerized society," *Sci. Eng. Ethics*, vol. 2, no. 1, pp. 25–42, Mar. 1996.
- [34] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *Int. Data Priv. Law*, vol. 7, no. 2, pp. 76–99, May 2017.
- [35] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *SSRN Electron. J.*, vol. 31, no. 2, pp. 841–887, 2018.
- [36] L. A. Bygrave, "Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling," *Comput. Law Securiry Rev.*, vol. 17, no. 1, pp. 17–24, 2001.
- [37] Ministerie van Financiën, *Regeling eed of belofte financiële sector 2015*, vol. BWBR0036152. 2015.
- [38] United Nations, "Universal Declaration of Human Rights," 12-Oct-1948. [Online]. Available: <https://www.un.org/en/universal-declaration-human-rights/index.html>. [Accessed: 08-Oct-2019].
- [39] IBM, "CRISP-DM Help Overview," 24-Oct-2014. [Online]. Available: undefined. [Accessed: 02-Jul-2019].
- [40] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, Manchester, UK, 2000, pp. 29–39.
- [41] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Aug. 2018.
- [42] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, San Francisco, California, USA, 2016, pp. 1135–1144.
- [43] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *ArXiv170507874 Cs Stat*, May 2017.
- [44] R. M. Grath et al., "Interpretable Credit Application Predictions With Counterfactual Explanations," Dec. 2018. In *NIPS 2018 workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy*, Dec 2018, Montreal, Canada.
- [45] D. Gunning, "Explainable Artificial Intelligence (XAI)." Nov-2017.
- [46] L. Floridi et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds Mach.*, vol. 28, no. 4, pp. 689–707, Dec. 2018.
- [47] B. Y. Lim and A. K. Dey, "Assessing Demand for Intelligibility in Context-aware Applications," in *Proceedings of the 11th International Conference on Ubiquitous Computing*, New York, NY, USA, 2009, pp. 195–204.
- [48] O. Biran and K. McKeown, "Justification narratives for individual classifications," in *AutoML Workshop*, Beijing, China, 2014, vol. 32, p. 7.
- [49] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? Ways explanations impact end users' mental models," in *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, San Jose, CA, USA, 2013, pp. 3–10.
- [50] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of Explanatory Debugging to Personalize Interactive Machine Learning," in *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, Atlanta, Georgia, USA, 2015, pp. 126–137.
- [51] J. Davis and L. P. Nathan, "Value Sensitive Design: Applications, Adaptations, and Critiques," in *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, J. van den Hoven, P. E. Vermaas, and I. van de Poel, Eds. Dordrecht: Springer Netherlands, 2015, pp. 11–40.
- [52] L. Mok and S. Hyysalo, "Designing for energy transition through Value Sensitive Design," *Des. Stud.*, vol. 54, pp. 162–183, Jan. 2018.
- [53] B. Taebi, A. Correljé, E. Cuppen, M. Dignum, and U. Pesch, "Responsible innovation as an endorsement of

- public values: the need for interdisciplinary research,” *J. Responsible Innov.*, vol. 1, no. 1, pp. 118–124, Jan. 2014.
- [54] V. Vaishnavi, W. Kuechler, and S. Petter, “Design Science Research in Information Systems.” 20-Dec-2017.
- [55] I. van de Poel, “Translating Values into Design Requirements,” in *Philosophy and Engineering: Reflections on Practice, Principles and Process*, vol. 15, D. P. Michelfelder, N. McCarthy, and D. E. Goldberg, Eds. Dordrecht: Springer Netherlands, 2013, pp. 253–266.
- [56] R. I. Faulconbridge and M. J. Ryan, *Systems Engineering Practice*. Canberra, Australia: Argos PRes, 2014.
- [57] C. L. Dym, P. Little, and E. J. Orwin, *Engineering design: a project-based introduction*, 4. ed. New York: Wiley, 2014.
- [58] D. Martens, J. Huysmans, R. Setiono, J. Vanthienen, and B. Baesens, “Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring,” in *Rule Extraction from Support Vector Machines*, J. Diederich, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 33–63.

Appendixes

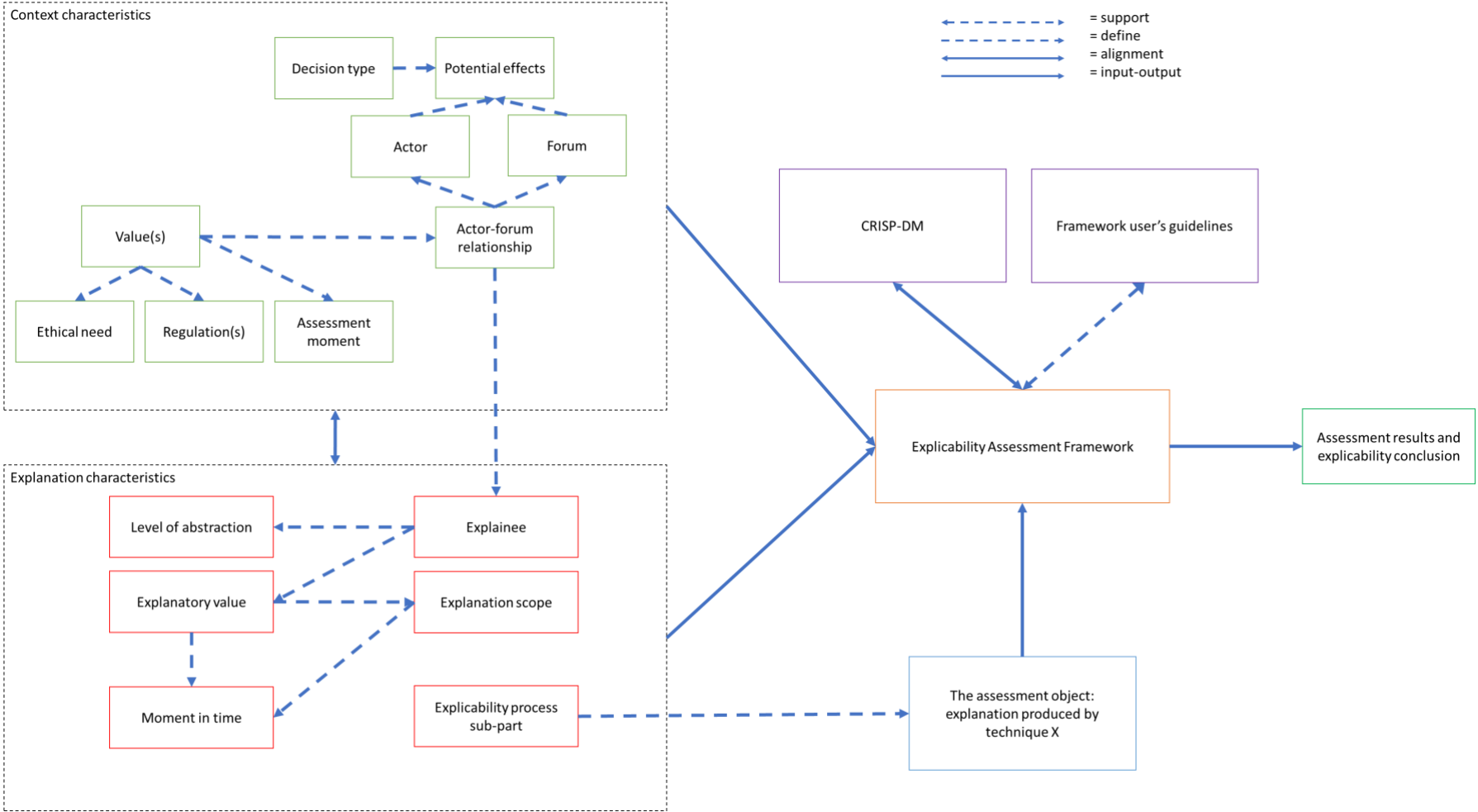
Appendix 1. Morphological chart

Function - Means	Means #1	Means #2
[FR.A1] shall show which new tasks need to be performed in the development lifecycle	User's guidelines that describe which tasks to perform	
[FR.A2] shall show how the new tasks need to be performed in the development lifecycle	User's guidelines that describe how the tasks need to be performed	
[FR.A3] shall show in what sequence the tasks need to be performed in the development lifecycle	User's guidelines that state in what sequence the tasks need to be performed	
[FR.A4] shall show at what times in the development lifecycle the tasks need to be performed	User's guidelines that show at what times in the CRISP-DM lifecycle the tasks are included	
[FR.B1] shall result in a conclusion if the explanation of interest is good enough	A conclusion section that synthesizes the assessments of the individual sections that concludes whether the explanation is good enough	
[FR.B2] shall give an overview of the specific explicability issue(s) in the system if it is not explicable enough	A summarizing section that shows the explicability issues to solve in the case of insufficient explicability and whereto the iteration step needs to go to accomplish this	
[C.C1] shall be usable for explanation assessment before the machine learning system is deployed	n/a - constraint	
[C.C2] shall be technique-agnostic, thus usable to assess different formats of explanations	n/a - constraint	
[FR.D1] shall be able to check which 'evidence roles' are useful to be included in the explanation	An overview of the evidence roles with examples and a usefulness check of these for the explanation	
[FR.D2] shall be able to check which 'evidence roles' are present in the explanation	Contains the question for all important evidence roles: "Does the explanation contain the *X-evidence role*?"	
[FR.E1] shall be able to specify the knowledge level or expertise that the explainee needs to have in order to understand the explanation	Contains the question: "What knowledge does the explainee need to have in order to understand the explanation?"	
[FR.E2] shall be able to conclude whether the explanation is understandable enough for a layperson	Contains the question: "Is the answer for FR.E1 limited enough to be considered understandable for a layperson?"	
[FR.F1] shall be usable to check which intelligibility types the explanation requires to fulfill a sufficient level of completeness	An overview of the intelligibility types and a usefulness check of these for the explanation	

Scientific paper manuscript of the graduation thesis by N.J. Herber

[FR.F2] shall be usable to assess the extent to which the relevant aspects are included in the explanation	Contains the question for the relevant intelligibility types: “Does the explanation address *intelligibility type X?”	
[FR.G1] shall be usable to assess the extent to which each component of an explanation’s content is truthful to how the underlying system took the decision	Contains the question: “Is the explanation a correct representation of how the model came to the decision; i.e. is the explanation based on the full truth, a simplified truth model, the truth of a singular feature or not the truth?”	
[FR.H1] shall be usable for the assessment of the narrative aspect of an explanation	Contains the question: “is the explanation textually written in a narrative format?”	Contains the question: “does the explanation include singular facts or datapoints without context?”
[FR.H2] shall be usable for the assessment of the human-interpretable linguistic aspects of the explanation	Contains the question: “Is the language used considered easily understandable for humans?”	Contains the question: “does the explanation sufficiently link the decision, important features, the roles and effects of these features in a logical way?”
[FR.I1] shall be able to assess the explanation length	Contains the question: “How many textual lines does the explanation include?”	Contains the question: “How many words does the explanation include?”
[FR.I2] shall be able to assess the number of concepts included in the explanation	Contains the question: “how many concepts are included in the explanation?”	
[FR.I3] shall be able to assess the modularity of the explanation structure	Contains the question: “Is the explanation structure considered modular; i.e. can the explanation easily be extended when consumers ask for additional explanation, without losing the structure of the explanation?”	

Appendix 2. Assessment framework relations



Appendix 3. Explicability Assessment Framework (EAF)

1. Context characteristics		
<i>Decision type</i>	*insert description of the type of decision that the machine learning system makes in the case of interest*	
<i>Value(s)</i>	*insert the value of interest for the design of the machine learning system in the use case*	
<i>Actor-forum relationship</i>	*insert a description of the type of relationship the actor and the forum have*	
<i>Forum</i>	*insert the forum of the decision*	
<i>Actor</i>	*insert the actor of the decision*	
<i>Potential effects</i>	*insert a description of the potential effects for the forum and actor of the decision*	
<i>Ethical need</i>	*insert a description of the ethical need that drives incorporating the value in the design of the system*	
<i>Regulation(s)</i>	*insert a description of the regulation(s) that drives incorporating the value in the design of the system*	
2. Explanation characteristics		
<i>Explainee</i>	Layperson/business/data-scientist/auditor	
<i>Level of abstraction</i>	Consumer-level/business-level/machine-level	
<i>Explanatory value</i>	Justification/teaching	
<i>Explanation scope</i>	Local/global	
<i>Moment in time</i>	Ex-post/ex-ante	
<i>Explicability process sub-part</i>	product/cognitive process/social process	
3. Framework adjustments		
<i>Evidence roles selection (justification)</i>		
	Normal evidence	Normal counter-evidence
	Exceptional evidence	Exceptional counter-evidence
	Contrarian evidence	Contrarian counter-evidence
	Missing evidence	Missing counter-evidence
<i>Intelligibility types selection</i>		
	Input	Output
	Why	How
	Why not	What if
	What else	Visualization
	Certainty	Control
	Situation	
4. Assessment Object Assessment		
insert the full explanation to be assessed		
<i>Questions:</i>		<i>Answers:</i>

<i>Justificatory explanation</i>	
1.X For evidence role X, does the explanation contain this evidence role?	
<i>Explanation towards a layperson</i>	
2.1 What knowledge does the explainee need to have in order to understand the explanation?	
2.2 Is the answer of 2.1 conform to the corresponding knowledge of the explainee?	
<i>Completeness</i>	
3.X For intelligibility type X, does the explanation contain this intelligibility type?	
<i>Soundness</i>	
4 Is the explanation a correct representation of how the model came to the decision; i.e. is the explanation based on the full truth model, a simplified truth model, the truth of (a) singular feature(s) or not on the truth?	
<i>Comprehensibility</i>	
5.1 Is the explanation textually written in a narrative format?	
5.2 Does the explanation include singular facts of datapoints without context?	
6.1 Is the used language considered easily understandable for humans?	
6.2 Does the explanation sufficiently link the decision, important features, the roles and the effects of these features in a logical way?	
<i>Conciseness</i>	
7.1 How many textual lines does the explanation include?	
7.2 How many words does the explanation include?	
8 How many concepts are included in the explanation?	
9 Is the explanation structure considered modular; i.e. can the explanation easily be extended when consumers ask for additional explanation, without losing the structure of the explanation?	
5. Concluding section	
*insert conclusion ... *	

Appendix 4. EAF application case study 1:

1. Context characteristics		
<i>Decision type</i>	The machine learning system decides if the credit applicant is classified as good or bad (i.e. a high vs low probability of default)	
<i>Value(s)</i>	Explicability	
<i>Actor-forum relationship</i>	Public accountability relationship between a bank and the credit applicant (consumer)	
<i>Forum</i>	Credit applicant	
<i>Actor</i>	Bank	
<i>Potential effects</i>	Receiving/being denied a loan, having more <i>good outstanding</i> loans/less <i>bad outstanding</i> loans for banks, risk of discrimination for credit applicants by banks, and as a consequence: legal prosecution, or publicly released news-item(s) that could cause damage to the brand, or distrust	
<i>Ethical need</i>	“Explanation capability towards the consumer is of crucial importance in a domain where the model needs to be validated before being implemented”	
<i>Regulation(s)</i>	N/a – no specific geographical area, but a more general use case	
2. Explanation characteristics		
<i>Explainee</i>	Layperson	
<i>Level of abstraction</i>	Consumer-level	
<i>Explanatory value</i>	Justification	
<i>Explanation scope</i>	Local	
<i>Moment in time</i>	Ex-post	
<i>Explicability process sub-part</i>	Explanation product	
3. Framework adjustments		
<i>Evidence roles selection (justification)</i>		
	Normal evidence	Normal counter-evidence
	Exceptional evidence	Exceptional counter-evidence
	Contrarian evidence	Contrarian counter-evidence
	Missing evidence	Missing counter-evidence
<i>Intelligibility types selection</i>		
	Input	Output
	Why	How
	Why not	What if
	What else	Visualization
	Certainty	Control
	Situation	
4. Assessment Object Assessment		
<p>Example rule set used to result in decision X (applicant is good/bad):</p> <p>“if (Checking Account < 0DM) and (Housing = rent)</p> <p>then Applicant = Bad</p> <p>elseif (Checking Account < 0DM) and (Property = car or other) and (Present residence since ≤ 3y)</p> <p>then Applicant = Bad</p> <p>elseif (Checking Account < 0DM) and (Duration ≥ 30m)</p> <p>then Applicant = Bad</p> <p>elseif (Credit history = None taken/All paid back duly)</p> <p>then Applicant = Bad</p>		

<p><i>elseif (0 ≤ Checking Account < 200DM) and (Age ≤ 28) and (Purpose = new car) then Applicant = Bad else Applicant = Good”</i></p>	
Questions:	Answers:
<i>Justificatory explanation</i>	
1.1 Does the explanation contain normal evidence?	No
1.2 Does the explanation contain normal counter-evidence?	Yes, the <i>low amount at checking account, long duration, age < 28</i> are features that are expected to influence the decision negatively (towards applicant = bad)
1.3 Does the explanation contain contrarian evidence?	No
1.4 Does the explanation contain contrarian counter-evidence?	Yes, <i>housing = rent, property = car, purpose = new car, credit history = none taken/all paid back duly</i> are not expected to directly influence the decision negatively (it is not directly clear why this is the case), but it does.
<i>Explanation towards a layperson</i>	
2.1 What knowledge does the explainee need to have in order to understand the explanation?	How an if-else statement works, the definitions of checking account, DM, Duration, Credit history, present residence
2.2 Is the answer of 2.1 conform to the corresponding knowledge of the explainee?	A layperson might not have the knowledge to understand how an if-else statement works, what DM means
<i>Completeness</i>	
3.1 Does the explanation contain this intelligibility type “why”?	No, but can easily be seen with regards to the input data of the applicant, of which it is aware
3.2 Does the explanation contain this intelligibility type “how”?	Yes, it shows how the decision has been made
3.3 Does the explanation contain this intelligibility type “why not”?	Yes, it shows why the other decision has not been made
<i>Soundness</i>	
4 Is the explanation a correct representation of how the model came to the decision; i.e. is the explanation based on the full truth model, a simplified truth model, the truth of (a) singular feature(s) or not on the truth?	It is a simplified truth model, derived from the Support Vector Machine used to make a decision
<i>Comprehensibility</i>	
5.1 Is the explanation textually written in a narrative format?	It is textually written, however, it is not in a narrative format
5.2 Does the explanation include singular facts of datapoints without context?	Yes, there is no context on the reasoning why certain rules are in place
6.1 Is the used language considered easily understandable for humans?	No, there are signs (</<=) and statements (elseif) that are not directly clear for humans what it means
6.2 Does the explanation sufficiently link the decision, important features, the roles and the effects of these features in a logical way?	Yes, it does link the decisions, the important features and their roles for the decisions in a logical if-else rule relationship
<i>Conciseness</i>	


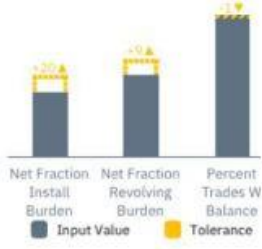

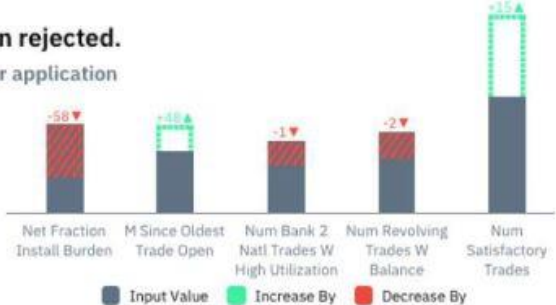
7.1 How many textual lines does the explanation include?	13
7.2 How many words does the explanation include?	84
8 How many concepts are included in the explanation?	9
9 Is the explanation structure considered modular; i.e. can the explanation easily be extended when consumers ask for additional explanation, without losing the structure of the explanation?	Yes, extra rules can easily be added.
5. Concluding section	
<p>The explanation is insufficiently good, and therefore the ML system within this context and specific explanation is insufficiently explicable, because:</p> <ul style="list-style-type: none"> - The explanation includes contrarian counter-evidence - A layperson requires knowledge to understand the explanation that it might not has - It misses a clear statement on why a certain decision has been made, although this can easily be derived from the decision tree - It misses the narrative format - It misses the context of why certain rules are there that influence the decision - It contains signs and statements that are not directly considered clear for a human what is meant with it <p>In short: it lacks the layperson perspective, it is not complete enough, it is not comprehensible enough.</p>	

For the designers of the machine learning system that has been used for the decision, this means that they have to do the following in a back-iteration to the business understanding step:

- First, they have to find out how it comes that contrarian counter-evidence is included in the explanation and has influenced the decision in an unexpected way. This could indicate a problem with the validity of the model, but this is not necessarily true. If there is a problem, the designers obviously have to solve this in order to be able to ensure the validity and the ability to justify the decisions.
- Second, the explanation should be enhanced with simpler language, that can be more easily understood and the unclear signs and statements must be replaced.
- Third, a why statement and context to the rules need to be added, after which the full explanation needs to be transformed into a narrative format

Appendix 5. EAF application case study 2:

1. Context characteristics		
Decision type	The machine learning system predicts a variable called ‘RiskPerformance’. “The value “Bad” indicates that a consumer was 90 days past due or worse at least once over a period of 24 months from when the credit account was opened. The value “Good” indicates that they have made their payments without ever being more than 90 days overdue.”	
Value(s)	Explicability	
Actor-forum relationship	Public accountability relationship between a bank and the credit applicant (consumer)	
Forum	Credit applicant	
Actor	Bank	
Potential effects	Receiving/being denied a loan, having more <i>good outstanding</i> loans/less <i>bad outstanding</i> loans for banks, risk of discrimination for credit applicants by banks, and as a consequence: legal prosecution, or publicly released news-item(s) that could cause damage to the brand, or distrust.	
Ethical need	“Explaining predictions of black-box models is of uttermost importance in the domain of credit risk assessment”	
Regulation(s)	“The problem is even more prominent given the recent right to explanation introduced by the European General Data Protection Regulation Goodman and Flaxman [2016], and a must due to regulation in the financial domain.”	
2. Explanation characteristics		
Explainee	Layperson	
Level of abstraction	Consumer-level	
Explanatory value	Justification	
Explanation scope	Local	
Moment in time	Ex-post	
Explicability process sub-part	Explanation product	
3. Framework adjustments		
Evidence roles selection (justification)		
	Normal evidence	Normal counter-evidence
	Exceptional evidence	Exceptional counter-evidence
	Contrarian evidence	Contrarian counter-evidence
	Missing evidence	Missing counter-evidence
Intelligibility types selection		
	Input	Output
	Why	How
	Why not	What if
	What else	Visualization
	Certainty	Control
	Situation	
Assessment Object Assessment		

<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 45%;">  <p>Congratulations, your loan application has been approved.</p> <p>If instead you had the following values, your application would have been rejected:</p> <ul style="list-style-type: none"> • NetFractionRevolvingBurden: 55 • NetFractionInstallBurden: 93 • PercentTradesWBalance: 68 </div> <div style="width: 45%; text-align: right;">  </div> </div> <p style="text-align: center;">(a) Positive counterfactual explanation</p>	
<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 45%;">  <p>Sorry, your loan application has been rejected.</p> <p>If instead you had the following values, your application would have been approved:</p> <ul style="list-style-type: none"> • MSinceOldestTradeOpen: 161 • NumSatisfactoryTrades: 36 • NetFractionInstallBurden: 38 • NumRevolvingTradesWBalance: 4 • NumBank2NatlTradesWHighUtilization: 2 </div> <div style="width: 45%; text-align: right;">  </div> </div> <p style="text-align: center;">(b) Counterfactual explanation</p>	
Questions:	Answers:
<i>Justificatory explanation</i>	
1.1 Does the explanation contain normal evidence?	Yes, <i>MSinceOldestTradeOpen</i> and <i>NumSatisfactoryTrades</i> are expected to influence the decision positively (lower probability for ‘bad’ target), and they do.
1.2 Does the explanation contain normal counter-evidence?	Yes, <i>NetFractionInstallBurden</i> and <i>NumBank2NatlTradesWHighUtilization</i> are expected to influence the decision negatively (higher probability for ‘bad’ target), and they do.
1.3 Does the explanation contain contrarian evidence?	No
1.4 Does the explanation contain contrarian counter-evidence?	Yes, <i>NumRevolvingTradesWBalance</i> is not expected to influence the decision strongly negative but it does.
<i>Explanation towards a layperson</i>	
2.1 What knowledge does the explainee need to have in order to understand the explanation?	The definitions and effects on the decision of: <i>NumRevolvingTradesWBalance</i> , <i>NetFractionInstallBurden</i> , <i>NumBank2NatlTradesWHighUtilization</i> , <i>NumSatisfactoryTrades</i> and <i>MSinceOldestTradeOpen</i> .
2.2 Is the answer of 2.1 conform to the corresponding knowledge of the explainee?	A layperson might not have the knowledge to understand what the definitions of the explaining features mean and what their effect is on the decision.
<i>Completeness</i>	
3.1 Does the explanation contain this intelligibility type “why”?	Yes, it shows why the decision has been made.
3.2 Does the explanation contain this intelligibility type “how”?	No, it does not show how the decision has been made.
3.3 Does the explanation contain this intelligibility type “why not”?	Yes, it shows why the other decision has not been made.

<i>Soundness</i>	
4 Is the explanation a correct representation of how the model came to the decision; i.e. is the explanation based on the full truth model, a simplified truth model, the truth of (a) singular feature(s) or not on the truth?	The explanation is based on the truth of a few singular features that the decision is based on.
<i>Comprehensibility</i>	
5.1 Is the explanation textually written in a narrative format?	It is textually written, with a supportive visualization, however, it is not in a narrative format.
5.2 Does the explanation include singular facts of datapoints without context?	Yes, the explanation contains features with singular values without context around these features or values.
6.1 Is the used language considered easily understandable for humans?	Yes, it contains an easily understandable enumeration of an answer to the main question: what values does the application need to be approved?
6.2 Does the explanation sufficiently link the decision, important features, the roles and the effects of these features in a logical way?	No, it does include the decision and the important features, however, it does not include a link of these in a logical way including their effects and roles
<i>Conciseness</i>	
7.1 How many textual lines does the explanation include?	7
7.2 How many words does the explanation include?	30
8 How many concepts are included in the explanation?	5
9 Is the explanation structure considered modular; i.e. can the explanation easily be extended when consumers ask for additional explanation, without losing the structure of the explanation?	Yes, additional features and/or additional explanation to the features can easily be added
Concluding section	
<p>The explanation is insufficiently good, and therefore the ML system within this context and specific explanation is insufficiently explicable, because:</p> <ul style="list-style-type: none"> - The explanation includes contrarian counter-evidence - A layperson requires knowledge to understand the explanation that it might not have - It misses a clear statement on how the decision has been made - It misses the narrative format - The explanation contains singular values without context - The explanation does not link the decision and important features with their effects and roles <p>In short: It lacks the layperson perspective, it is not complete enough, it is not comprehensible enough.</p>	

For the designers of the machine learning system, this means that they have to do the following in a back-iteration to the business understanding step:

- First, again the reason for the contrarian counter-evidence need to be found in order to find out whether this indicates a problem with the validity of the model and ability for decision justification.
- Second, the explanation should be improved: simpler language, that can be more easily understood, and links between the decision and important features, effects and roles should be added.

Third, a how-statement and context to the singular values need to be added, after which the full explanation needs to be transformed into a narrative format.